ESTIMATION OF TWO-PARAMETER MULTILEVEL ITEM RESPONSE MODELS WITH PREDICTOR VARIABLES: SIMULATION AND SUBSTANTIATION FOR AN URBAN SCHOOL DISTRICT

A Dissertation

by

PRATHIBA NATESAN

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2007

Major Subject: Curriculum and Instruction

ESTIMATION OF TWO-PARAMETER MULTILEVEL ITEM RESPONSE MODELS WITH PREDICTOR VARIABLES: SIMULATION AND SUBSTANTIATION FOR AN URBAN SCHOOL DISTRICT

A Dissertation

by

PRATHIBA NATESAN

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Co-Chairs of Committee,	Norvella P. Carter
	Bruce Thompson
Committee Members,	Stephanie L. Knight
	Dennie L. Smith
Head of Department,	Dennie L. Smith

August 2007

Major Subject: Curriculum and Instruction

ABSTRACT

Estimation of Two-Parameter Multilevel Item Response Models with Predictor Variables: Simulation and Substantiation for an Urban School District. (August 2007)

Prathiba Natesan, B. Arch., University of Madras;

M.S., Texas A&M University

Co-Chairs of Advisory Committee: Dr. Norvella P. Carter
Dr. Bruce Thompson

The most recent development in the field of Item Response Theory (IRT) has been the evaluation of IRT models as multilevel models, known as Multilevel IRT models (MLIRT). These models offer several statistical and practical advantages over ordinary IRT models. However, models such as 2-PL MLIRT models have not been studied yet. This dissertation consists of two studies, a simulation and a substantiation for an urban school district dataset. The simulation study tested the performance of two-parameter (2-PL) MLIRT models with predictor variables under various conditions that included 3 test lengths (15, 30, and 60 items), 4 sample sizes (200, 500, 1000, and 2000), 2 correlation conditions between the predictor variable and the ability (or attitude) parameter (r_{pb}=.35 and .8), and 4 binomial distributions of the predictor variable (p=0.1, 0.25, 0.4, and 0.5).

The bias and Root Mean Square Deviation (RMSD) values of the item parameters indicated that the distribution of the predictor variable and the correlation

between the predictor and the ability (or attitude) parameter did not affect the estimates of 2-PL MLIRT models. These models performed well for sample sizes as low as 500 and test lengths as low as 15 which is lower than the required sample size for ordinary IRT models. Even for a sample size of 200, sufficiently accurate estimates were obtained with more than 300 iterations.

The second study investigated the characteristics of the items that measured urban teachers' perceptions of cultural awareness and beliefs about teaching African American children and tested whether these perceptions were influenced by the teachers' gender, ethnicity, or teaching experience. Teacher beliefs about teaching African American students, culturally responsive management, and cultural awareness factors were influenced by the ethnicity of the teachers. Culturally responsive management, home and community support, and curriculum and instructional strategies factors were influenced by the teaching experience of the teachers. Items that were biased based on ethnicity or teaching experience were identified. None of the items exhibited gender bias. The study identified items that could be used over other items when the need for a shorter instrument or more informative categories arises.

DEDICATION

TO AMMA AND KANNA

ACKNOWLEDGEMENTS

I begin my acknowledgements by thanking my amma (mother), who has taught me the value of being strong, fighting against all odds to achieve one's dream, and above all, the importance of education. She is the bravest individual I know and the only god I have seen. Kannan (my brother), who holds the record for both getting beaten by me the most and, in turn, beating me the most, has continued to be my source of inspiration to question everything and think critically. Our bond grows stronger every day as we share the bitter and the sweetest moments of our lives. Last, but not the least, in my weird family is my (really) better half, Praveen, who has guided, helped, and argued with me so much that his contribution to my achievements is inevitable. My life would mean "zilch" without any of these people.

I cannot imagine doing my PhD without the unconditional support of Dr. Carter who has been my friend, philosopher, and guide. She has seen me through the good, the bad, and the ugly and has always been there for me. My gratitude to her cannot simply be expressed in words! Thank you, Dr. Carter for taking me under your wing and being a mother to a girl away from home.

Bruce, thank you. I fell in love with statistics because of you. You have been a wonderful mentor, teacher, and guide for me throughout my doctoral education. It always amazes me that someone who looks so distant can be so caring when it comes to the professional life of students. Thank you, Dr. Thompson for just being who you are and teaching me so much.

I would also like to thank Dr. Smith and Dr. Knight for providing me with support and counseling, both professionally and academically, when needed. I have never seen a department head who is so concerned about the well being of his/her students. My thanks to Dr. Kracht and Erin for making us a part of their lives. Your house will always be a home away from home! My discussions and sessions with Dr. Knight, Dr. Varni, and Dr. Greenwald have helped me develop intellectually. I would also like to thank Dr. Webb-Johnson for her permission to use the CABI instrument, Dr. Young-Hawkins for her constant encouragement, and Dr. Willson for his guidance on several statistical topics.

Soma's contribution to this dissertation saved my life from being a constantly-fearing nightmare. Your statistical genius and "communication" style will never be forgotten. My gratitude to L.C. who has always believed in me and taught me that while I might not be as intelligent as I want to be, I was not as stupid as I thought myself to be. My thanks will be incomplete without acknowledging the support given by my friends, Aditi, Bhoo, Preetha, Jigyasa, Bill (Carter), China, Jim (Varni), Mike (Greenwald), Vidya, Samhita, both the Meeras, Miriam, Christine, Prashanti, Pudhuma, Walt, and my ex-roommates, Lakshmi, Myti, and Bhavani. I thank my students who have constantly improved my knowledge through questioning. Finally, I thank the music and rhythm in me that make my life worth living.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

CHAPTER I

INTRODUCTION

Item Response Theory (IRT) or Latent Trait Theory (LTT) has been widely used in educational testing (e.g., Bielinski & Davison, 2001; Edward, 1993; Kulick & Hu, 1989; Le, 1999; Pike, 1990; Rock, Pollack, & Quinn, 1995; Zwick & Ercikan, 1989) and more recently in health-related research (Hays, Morales, & Reise, 2000). However, various other research applications of this extremely useful method, especially in the field of education, have not been fully explored. The very name, latent trait theory, indicates this theory can be used to measure the unobservable traits of people through estimation and analysis of latent variables.

Therefore, ideally, during the past half a century when IRT was being developed and extensively utilized for educational testing, it should have been used in measuring diverse traits of people, such as the extent of their cultural awareness, teacher preparation, or even the extent of parental involvement in their children's academic progress. Additionally, IRT could be used to understand the characteristics of items that measure various traits or attitudes. This can further help in more efficient administration of instruments. A systematic literature review of the recent most 108 articles published in peer reviewed journals that use IRT shows otherwise as shown in Chapter II. The present research presents both the technical aspects of statistical development and the substantive application of IRT in everyday research.

This dissertation follows the style of *Educational and Psychological Measurement.*

Several models ranging from 1-parameter (1-PL) to 3-parameter (3-PL) models (sometimes even 4-PL models (Reise & Waller, 2003)), in addition to methods and computer programs to estimate these models, have been developed over the years. Newer developments in IRT include the multilevel IRT (MLIRT) model. This method provides many advantages such as including other predictor variables in the model, estimating item difficulty and person ability parameters simultaneously, and including group membership in the model. However, two-parameter (2-PL) MLIRT models have not been widely used.

This dissertation consists of two studies, the first of which explores the performance of two-parameter multilevel IRT (MLIRT) models for various test lengths, sample sizes, correlation between the predictor variable and the ability parameter, and the distributions of the predictor variable through simulation studies. The simulation of data was performed using MATLAB 7.1 and the parameters were estimated using WINBUGS 14. The second study was a substantive study applying the 2-PL MLIRT model to teachers' perceptions of Cultural Awareness and Beliefs Inventory (CABI) data to understand the characteristics of the items that measured teachers' cultural awareness. In doing so, the effects of variables such as gender, ethnicity, and educational level on cultural beliefs were assessed. The overall and differential effects of the covariates were estimated using Generalized Linear Latent and Mixed Models (GLLAMM) programs written using STATA by Rabe-Hesketh, Skrondal, and Pickles (2004a; 2004b).

Purpose of the Study

The purpose of this dissertation is to estimate the accuracy and precision of the parameters of the 2-PL MLIRT models for varying test lengths, sample sizes, correlation between the predictor variable and the ability parameter, and distributions of the predictor variables. This research compared estimates of item and person parameters from the 2-PL MLIRT model and the generated values of the parameters for various samples generated through Monte Carlo simulation for varying test lengths (15, 30, and 60 items), sample sizes (200, 500, 1000, and 2000), correlation between the predictor variable and the ability parameter (0.3 and 0.85), and distributions of the predictor variable (p=0.1, 0.25, 0.4, and 0.5). Thus, the minimum test lengths, sample sizes, correlation between the predictor variable and the ability parameter, and predictor variable distribution for which the 2-PL MLIRT model would produce sufficiently accurate estimates were identified. Further, the 2-PL MLIRT model was applied to the teachers' perceptions of Cultural Awareness and Beliefs Inventory (CABI) to understand the items that measured the perceptions of cultural awareness and beliefs of teachers in urban areas. Thus, the study afforded a comparison between the results of simulation studies and substantive studies.

Problem Statement

There is a statistical need to understand the 2-PL MLIRT model with respect to the minimum test lengths, sample sizes, and the type of predictors that would be required to produce sufficiently accurate estimates of model parameters. Although the 1-PL or the Rasch model has been emphasized in the field of IRT, the discrimination parameter is an

extremely important index which indicates the item-test correlation and the response consistency of the item (Reise, 2000). There is also a need to apply IRT in other fields of educational measurement. Such applications help understand the characteristics of the items and also identify the differential item functioning or the bias in items when measuring a certain trait across groups. Thus, applying the 2-PL MLIRT model to the CABI explores the characteristics of items that measure cultural awareness of teachers and also identifies the effect of ethnicity, gender, age, and years of teaching on the cultural awareness of teachers. These characteristics include the amount of information they contribute to the scale (which is useful when a concise form of the instrument has to be constructed) and the function of the different categories in the items (which is useful in building categories that could measure the attitudes more accurately).

Research Questions

1a. What is the accuracy and precision of the item parameter estimates of the 2-PL MLIRT model for datasets with varying test lengths (15, 30, and 60 items)?

1b. What is the accuracy and precision of the item parameter estimates of the 2-PL MLIRT model for datasets with varying sample sizes (200, 500, 1000, and 2000 examinees)?

1c. What is the accuracy and precision of the item parameter estimates of the 2-PL MLIRT model for datasets with varying correlations between abilities and the binomial predictor variable (r=0.3 and 0.85)?

1d. What is the accuracy and precision of the item parameter estimates of the 2-PL MLIRT model for datasets with normally distributed and extremely skewed predictor variables (p=0.1, 0.25, 0.4, and 0.5)?

2. How do test length, sample size, correlation between the predictor variable and the ability parameter, and distribution shape of the predictor variables interact to impact the accuracy and precision of the item parameter estimates of the 2-PL MLIRT model?

3. What are the characteristics of the items that measure teachers' perceptions of cultural awareness and as shown by the IRT analysis using the 2-PL MLIRT model?

4. Do these item characteristics of cultural awareness differ by gender, ethnicity, age, and the years of teaching of the teachers?

Limitations

1. The first limitation stems from the assumption that the item response functions will increase monotonically with increase in ability, albeit in real life some items have nonmonotonic functions (Levine, 1984; Samejima, 1979). Therefore there is less than exact replication of a real life scenario in the simulated data. This is a limitation for the simulation study.

2. The models simulated in this study are unidimensional IRT models. The limitation in simulating a unidimensional IRT model is that most of the latent traits are multifaceted and therefore governed by many factors although some simulation

studies have showed the difference between the unidimensional and multidimensional models to be trivial (e.g., Davey, Nering, & Thompson, 1997).

3. Usually in real life situations, tests are constructed very systematically. Therefore, when the item parameters are simply randomly selected from a distribution, the results tend to have decreased validity (Davey, Nering, & Thompson, 1997).

Statistical Need for the Study

Many applications of MLIRT have been published within the last decade (e.g., Adams, Wilson & Wu, 1997; Beretvas & Kamata, 2005; Beretvas, Meyers, & Rodriguez, 2005; Beretvas & Williams, 2004; Fox, 2004, 2005; Fox & Glas, 2001, 2003; Kamata, 1998, 2001; Maier, 2001; Raudenbush & Bryk, 2002; Raudenbush, Bryk, Cheong & Congdon, 2004). Based on the research of Adams, Wilson and Wu (1997), Mislevy (1985), and Raudenbush and Sampson (1999), Kamata (1998) developed 1-parameter multilevel IRT model estimation using HGLM for dichotomous data. This has further been extended for polytomous data (Williams, 2003) and for cross-classified data (Beretvas, Meyers, & Rodriguez, 2005).

However, 1-PL model only estimates the item difficulty which characterizes the position of the item on the ICC. The slope of the curve denoted by item discrimination *(a)*, which "indicates the quality or value of an item in the basic sense of the amount of information the item provides about θ [latent trait]" (Lord & Novick, 1968, p. 367), is not estimated. The discrimination parameter is especially important when it comes to adaptive testing which is an increasingly popular method of administering questionnaires

because items are tailored to the individual's abilities. The higher the item-test correlation, the higher is the discrimination parameter. Item discriminating power denotes the effectiveness of an item to discriminate among "poor" and "good" examinees. The more discriminating the items, the less the range of item difficulty, and hence a steeper curve (Lord & Novick, 1968). While the guessing parameter can be related to the trait level, and the item difficulty can be related to the endorsement rate of the item by an examinee (or where the probability of endorsement is 0.5), the discrimination parameter is related to the degree of response consistency of the item (Reise, 2000).

Because MLIRT involves extensive computations, an effective computer program is necessary to estimate the parameters and the model fit. Although several programs such as GLIMMIX (in SAS), MLwiN, HLM, NLMIXED, MIXOR/MIXNO, GLLAMM exist, in addition to specialized IRT softwares such as MULTILOG, BILOG, PARSCALE, each of these programs suffer from some limitations. Tuerlinckx et al. (2004) analyzed a dataset using GLIMMIX, GLLAMM, MLwiN, HLM, S-Plus, NLMIXED, and MIXOR/MIXNO and compared the estimates obtained from each of these programs. They found that the estimates were similar for identical models with downward biases for some programs (such as MLwiN, HLM, and GLIMMIX). However, because this was a one-shot study and not a simulation study, the results were not generalizable. Similarly, WINBUGS, a general purpose statistical software, offers a pre-written code that can be modified and used to perform IRT analysis. This leads us to the crux of the present research which is the need to conduct simulation studies to

understand the advantages and limitations of these programs under various conditions, such as different sample sizes, test lengths, and distributions.

Conceptual Need for the Study

Understanding the needs of diverse learners is one of the foremost challenges for teachers. According to many scholars (Gay 2000; Howard 2001; Ladson-Billings, 1994; Love, 2001; Villegas & Lucas, 2002) teachers' knowledge and implementation of culturally responsive pedagogy can impact and enhance the academic performance of students of color. However, for effective implementation of culturally responsive pedagogy, it is necessary to first know and understand the perceptions of teachers on cultural awareness and beliefs. Webb-Johnson and Carter (2005) developed an instrument that examined the cultural awareness and beliefs of urban teachers in order to develop intervention programs to help teachers with their pedagogical practices and to help narrow the gap between the learning styles of diverse learners and the teaching styles of teachers.

Most of the current research has used classical test theory and therefore, although the real intent would be to measure the latent trait of the person (such as attitudes, perceptions, quality), researchers usually tend to restrict the analysis to hypothesis testing or measuring relationship between or among measured variables. In such cases, the hypotheses would simply compare if the sample under study performs better than another group. In short, the aim of measuring the trait is not accomplished. CTT does not give an exact quantification of the trait itself. IRT, on the other hand, can give an exact measure of the trait of the person. However, before estimating the attitudes of teachers,

the characteristics of the items have to be studied. These item parameters would translate into the contribution of each item to the awareness of the teacher. Items with higher threshold or location parameters indicate lower endorsement rates by teachers. Similarly, items with higher discrimination parameters indicate that these items contribute more information towards the scale and also that the respondents are able to distinguish between the categories better on these items. The impact of each item on the cultural awareness of a teacher is different, unlike in CTT, where each item is given equal weightage or importance. Thus, by giving each item a differential share in contributing to the cultural beliefs of teachers, an effective measure of traits can be obtained. Moreover, the relationship between the items can also be studied, by studying the item parameters.

Significance of the Study

Conceptually, the present study is useful for researchers in several fields and not simply researchers in testing because it demonstrates the use of IRT models for estimating the latent trait, a commonly measured construct. Although the model has been widely known and other disciplines such as health, business, and marketing have adopted IRT into their mainstream research educational research has still confined IRT to mainly testing. Hopefully, the present study will increase the awareness of the applications of the IRT model in fields other than educational testing.

Statistically, through simulation this study investigated the minimum conditions necessary to estimate the parameters of a 2-PL MLIRT model in the presence of a predictor variable with adequate accuracy and precision. There are a variety of software

products available to perform IRT such as BIGSTEPS/WINSTEPS (Wright & Linacre, 1997), MULTILOG (Thissen, Chen, & Bock, 2003), and PARSCALE (Muraki & Bock, 1998). However, as Hays, Morales, and Reise (2000, p. 39) succinctly stated:

> None of these programs are particularly easy to learn and implement. The documentation is often difficult to read, and finding out the reason for program failures can be time consuming and frustrating. The existing programs have a striking similarity to the early versions of the LISREL structural equation-modeling program. LISREL required a translation of familiar equation language into matrixes and Greek letters. Widespread adoption of IRT … will be facilitated by the development of user-friendly software.

Both WINBUGS and GLLAMM are comparatively simpler than some of these programs to implement and estimate IRT models, and specifically, multilevel IRT models. The WINBUGS code can be modified according to user's needs to include several advanced methods of estimation such as Bayesian estimation. WINBUGS is a general purpose statistical software, usually used in medical research and is available free of cost over the internet. The pilot testing of the simulation was conducted both using WINBUGS and GLLAMM. However, due to time constraints and availability of more options, especially Bayesian estimation, WINBUGS was chosen over GLLAMM.

GLLAMM was used for the CABI study. GLLAMM can be recoded to use for other general use statistical software such as SAS and SPSS. Although the procedure is time-intensive GLLAMM does not require extensive programming knowledge but only takes longer processing time to process to produce estimates. By applying GLLAMM to the CABI data, this research demonstrated the application of IRT to data that are not necessarily test data thereby opening the door to wider applications of IRT. The

conceptual implications of this research may propagate the applications of IRT in both the quantitative and the qualitative worlds of research. Another hidden purpose of the two studies is to demonstrate how general purpose statistical software programs can be modified so that they can be used for IRT analysis.

Organization of the Study

This dissertation consisted of two studies, a simulation and a substantiation for an urban data set. Such a twin-study can help compare the results of a statistically perfect dataset that has been simulated and a real life dataset so that the applicability of the model in practical situations also can be evaluated. For the purposes of brevity and publication, the chapters in this dissertation were arranged in the form of journal articles. The first chapter presented the introduction to both the simulation and the substantiation studies. The second chapter is an article that consisted of a systematic literature review that demonstrates the need for the use of IRT in the measurement of diverse latent traits in educational research. The third chapter is an article that discusses the findings of the simulation study. The fourth chapter is an article that discusses the findings of the study for the urban district data for all the factors. The fifth chapter contains a brief summary and discussion that integrate the findings of both these studies and also discusses their implications and the scope for further research.

CHAPTER II

THE CASE FOR IRT IN LATENT TRAIT MEASUREMENT: A LITERATURE REVIEW

Introduction

Item Response Theory (IRT) or Latent Trait Measurement models were heralded as "one of the most important methodological advances in psychological measurement in the past half-century" (McKinley & Mills, 1989, p. 71). Their comparison to Classical Test Theory (CTT) is inevitable because CTT has been the most widely used measurement model to date. One limitation of CTT is its inability to separate the test characteristics from the examinee characteristics (Henard, 2000). In fact, Hambleton and Swaminathan (1985) said that within CTT it is difficult to say whether an item is easy or difficult, because in CTT this depends on the abilities of the examinees. Conversely, it is also difficult to say if an examinee is smart or not, because this depends on the difficulty level of the item being administered. Furthermore, in CTT, item difficulties and person abilities are on different scales (Wright & Stone, 1979).

IRT overcomes these limitations and helps the researcher build items free from examinee and test item biases (Henard, 2000; Wright & Stone, 1979). This theory transforms the item difficulties and person abilities into estimates on a single scale which is theoretically both "person-free" and "item-free", thereby taking care of the redundancy of cyclical dependence (Cantrell, 1999; Thompson, 2006). Furthermore, unlike CTT, in IRT, the relationship between the latent construct and the true score is nonlinear. In fact, in IRT, the probability of answering an item correctly is a logistic function (Raju, Laffitte, & Byrne, 2002).

Purpose of the Paper

Although IRT was developed by educational researchers and has been used the longest and the most in education, wider applications of IRT have not been explored. In educational research, IRT still remains confined to educational testing and academic achievement measurement. The purpose of this article is to present evidence of the dearth of IRT applications other than testing in education research and also suggests some directions for future research. In so doing, I also hope to instill the fact that IRT is not the 'enigma' that several researchers make it out to be, nor does IRT have only a narrow range of applications. If properly used, IRT can fill many of the voids in the current knowledge base. Therefore the main purpose of this article is to advocate the use of IRT for a wider range of applications.

IRT Postulates

IRT has two main postulates, the latent traits and the Item Characteristic Curve (ICC) (Cantrell, 1999). Latent traits (or simply traits or abilities) measure the performance

of an examinee on a test item. An ICC is a frequency polygon or an ogive representing the relationship between the item performance and the examinee's set of traits that determine the performance (Cantrell, 1999; Hambleton & Swaminathan, 1985). The ICC reflects the probability of selecting a certain response to an item with respect to the ability, attitude, or latent trait of the person (Ostini & Nering, 2006). There are three parameters, (a) the item discrimination parameter, "a", (b) the item difficulty parameter, "b", and (c) the guessing parameter, "c" (Lord & Novick, 1968), that determine the likelihood of an item being answered correctly and therefore the ICC.

These three parameters combined with the person's ability, attitude, or latent trait give rise to the ICC. While many scholars argue about the importance, the inadequacy, or the redundancy of these parameters, each parameter represents a different aspect of the probability of a given response. The guessing parameter can be related to the trait level and the item difficulty can be related to the endorsement rate of the item by an examinee (or where the probability of endorsement is 0.5) (Reise, 2000).

There are several IRT models based on the parameters used. While a three-parameter model consists of all these three parameters, a two-parameter model consists of only item discrimination and item difficulty parameters, omitting the guessing parameter. Therefore, a two-parameter model is the special case of the three-parameter model with guessing parameter always set to zero. Similarly, in a one-parameter model, known as the Rasch model (Rasch, 1960), both guessing and item discrimination parameters are not included and hence not modeled (Hambleton & Swamination, 1985). Therefore, a one-parameter model is a special case of the three-parameter model where the guessing

parameter is zero and all the slopes or discrimination parameters are equal and hence only the influence of item difficulty is considered.

In order to convert the ICCs that are not generally linear into approximately linear curves, "logits" or natural logarithmic values of the functions of proportions are used (Cantrell, 1999; Wright & Stone, 1979). The natural log of 1 is 0, 2 is 0.693, 3 is 1.098 and so on. It is clear that this conversion is not linear. Similarly, the curvilinear ICC gets converted to an almost linear "Logit" graph. Thompson (2006) explained in detail how logistic transformation of data makes curvilinear dynamics linear. Person ability, attitude, or latent trait will be estimated along with item difficulty as a latent trait simultaneously, so that the problem of their interdependency can be eliminated (Mislevy, 1987; Patz & Junker, 1999a, 1999b).

Is IRT Really Necessary? Both Sides of the Equation

Although IRT has proved to be a useful tool and several studies now increasingly use these models for data analysis, there have been critics of IRT models as well (e.g., Burton, 2004, 2005; Fan, 1998; Lawson, 1991; MacDonald & Paunonen, 2002). Burton (2004) pointed out that while performing item analysis, items are rejected when they do not fit the statistical model and this leads to not testing the knowledge of people in that particular area. Furthermore, test scores of zero or 100% are excluded from the analysis, because researchers are more concerned with "quantifying psychological traits than with testing knowledge" (Burton, 2004, p. 339) and IRT models cannot estimate the abilities of persons with these scores. The facts that IRT requires a sample size in the order of hundreds or sometimes thousands and involves complex computations do not make its

usage any simpler. Therefore, weighing all the advantages the method provides over CTT, some scholars claim that the effectiveness of IRT over CTT is limited (Burton, 2004, 2005; Lawson, 1991; MacDonald & Paunonen, 2002).

In spite of these disadvantages, IRT offers many advantages over CTT. One such advantage is that, unlike, CTT, IRT does not require item error variance to be equal across populations (Raju, Laffitte, & Byrne, 2002). In most practical situations it is almost impossible to find items with equal error variances which makes the results of classical test analyses much less robust (Byrne, 1994a, 1994b, 1998, 2001). In IRT, the error variance differs as a function of the person ability, attitude, or latent trait and therefore, can vary from person to person (Hambleton, Swaminathan, & Rogers, 1991).

Wider IRT Applications

The earlier sections discussed the merits of IRT and how its applications have come to be widely used in many fields of research. This section presents a systematic study of the 108 most recently published peer-reviewed journal articles that used IRT. These findings will reveal the myriad of information and research potential IRT offers in various aspects of measurement.

The literature search was restricted to keyword, abstract, or titles containing the word “Item Response Theory” on the *Worldcat* database. These articles were divided (a) subject-wise into education, statistics, health, business, other behavioral sciences and psychology-related, and both education and health-related; and (b) research application-wise into empirical, practical, empirical research with a practical example for illustrating the model, research perspectives, literature review, and book review. The subjects of the

articles were also coded in order to give an idea of the spectrum of research interests covered by these articles.

Of the 108 articles analyzed, 53 were related to or published in the field of education, 35 in health, 7 in statistics, 6 in business, and 5 in other behavioral sciences. Thirty three of these articles performed empirical research which consisted of research performed on simulated data sets to test or build new models or the conditions for robustness and the like. Forty four of these articles performed practical research that involved application of IRT to real life data such as the Headache Impact Test or the Depression Inventory. Seven articles were authors' perspectives on IRT, 3 were literature reviews, and one a book review. Interestingly, there were 18 articles which were empirical and used a practical dataset to explain the IRT model of interest. The individual division of the field that conducted the specific type of research can be seen in Table 1.

It is interesting to note that the field of education has contributed 87.88% of the empirical research articles to the knowledge base while also contributing 50% of the empirical/practical research studies. These studies covered a vast range of topics including estimation methods, validity, reliability, comparison with factor analysis or regression, dimensionality, multidimensional IRT, testlet pool construction, rater effect comparison, test equating, person fit, curve fitting, missing data, item dependence, sequential tests, Computer Adaptive Testing (CAT), and DIF. Such empirical studies are important to the development and testing of models and their conditions of robustness in order to apply IRT models to practical research appropriately. The field of education

contributes to the development of newer models more than any other field including statistics.

Table 1: Types of IRT Research Publications in Various Disciplines

Field of Study	Empirical	Practical	Perspectives	Lit. Review	Emp. w/ Prac. Illustration	*Total*
Education	29	12	3	0	9	***53***
Health	0	22	3	3	6	***34***
Business	0	4	1	0	2	***7***
Behavioral/ Psychology-related	0	6	0	0	0	***6***
Statistics	4	0	0	0	1	***5***
Total	***33***	***44***	***7***	***3***	***18***	***105****

** One article was a book review and therefore not included in the analysis*

The field of health, however, makes the maximum contribution to the practical applications of IRT and at the same time contributes one-third of the empirical/practical research articles to the knowledge base. This is an interesting finding because, although the empirical development of IRT takes place in the field of education more than any other, the application of IRT to mainstream practical research problems remains

comparatively unattended to in education. In fact, of the 11 practical application papers in education, 8 of them applied IRT to testing data. Only the other 3 papers involved measurement of attitudes, cultural equivalence, and leadership practices, which form a small part of the realm of latent traits that can be measured.

It is clear that when it comes to latent trait measurement, many educational research papers restrict themselves to testing hypotheses that involve comparison between/among groups, regression, or multilevel modeling to study the effect of one variable on the other. However, the latent trait itself is not measured. Most practical applications of IRT in education were confined to test equating, measurement of the effect of stimulus on test performance of examinees, validity assessment of items in large scale assessment, and assessment of student proficiency and dimensionality.

Consider a research that studies the verve level of students with an instrument that has items measuring different aspects of “verve”. A researcher would normally perform statistical significance testing between students who differed by ethnicity or gender or try to find the relationship between the different items such as the impact of verve on academic achievement. However, if IRT were to be used on this instrument, one could find the actual trait of focus here, verve, and then perform the same analyses using the IRT estimates of the latent trait.

Conclusion

There seems to be a dearth in the usage of IRT in the field of education when it comes to latent trait measurement other than academic achievement testing. While many powerful, empirical studies are being conducted, to develop newer and more sophisticated models and methods, these methods are only as useful as their applications. The benefits of IRT can achieve fruition only when these models see the light of practical applications. Other fields such as health care and business apply these models to real life data, but education lags far behind in this respect. Although, in the current scenario, academic achievement and achievement gap are the most crucial topics of interest in education, there remain many other questions that also require the sophistication and the options offered by IRT.

Of course, one should always remember that the method follows the question and also consider whether IRT's benefits outweigh IRT's complexities enough to perform IRT analyses because it is both complex and time-intensive. While IRT is a powerful tool to measure latent traits, there are large sample sizes that are required to perform IRT analysis. Moreover, the math behind IRT is complex and the available texts not simple enough which has led to several researchers believing IRT to be an ill-explained and unclear concept (e.g., Burton, 2004, 2005).

Therefore, considering both sides of the argument, one can conclude that although IRT cannot be used for all types of analyses and datasets, IRT is an extremely powerful tool that can fill the voids of the knowledge base by answering some questions that CTT cannot answer. However, there still remains a wide gap between developing

models in IRT and applying them to real life datasets, especially to measure some of the most commonly researched traits in the field of education. So, asking the earlier question again, “Should IRT always translate into educational testing?” the answer clearly is *no*. The issue of whether this idea will lure education researchers to utilize IRT for wider applications remains a question. If wider applications of IRT are studied, many more dimensions of the concept can be understood. The so-called “mysteries” of IRT would probably unravel as more studies with wider applications were performed. The hurdles along the path cannot be overcome unless they are first encountered. A research phenomenon can only be as useful as its application. It is time we gave IRT its due in the field of research.

CHAPTER III

TWO PARAMETER MULTILEVEL ITEM RESPONSE THEORY MODELS: A MONTE CARLO SIMULATION STUDY FOR TEST LENGTHS, SAMPLE SIZES, AND PREDICTOR VARIABLES

Multilevel IRT Models

The most recent development in the field of Item Response Theory (IRT) is the concept of Multilevel Item Response Theory Models (MLIRT). When the effects of multilevel covariates on a latent trait need to be estimated, IRT and Multilevel Modeling can be combined. This amalgamation of the two models allows us investigate and analyze the covariates that affect the person abilities instead of simply estimating the latent traits (Maier, 2001). This merger also paves way to modeling the abilities over time when repeated observations are made, or across various raters, or simply for people belonging to a certain group versus another.

The simplest way to combine the two methods is to consider items as nested within people (Adams, Wilson & Wu, 1997; Kamata, 1998). This facilitates the modeling of measurement error within and between these two levels. The traditional method of finding the effect of covariates on the person traits, by estimating the parameters which then form a part of the MLM, gives biased parameter estimates and this bias increases with decrease in sample size. Thus, MLIRT models are obtained by considering the item difficulty (location) as the first level variable and the person ability, attitude, or latent trait as the second level variable. In other words, the first level is an item level model and the second is the person level model. For a dichotomous variable, a

Bernoulli sampling model is used. Bernoulli sampling is the probability distribution in which each trial has two possible outcomes, success or failure, which in this case would be item correct or item incorrect. The trials have to be independent, which means the items are independent of each other and the probability of success is p for each trial and of failure is (1-p). For item i and person j,

$$Y_{ij} = \beta_{oj} + \beta_{1j} X_{1ij} + \ldots + \beta_{kj} X_{kj} \quad (1)$$

$$= \eta_{ij}$$

where Y_{ij} is the item response, β_{1j} is the effect of item 1 or the co-efficient associated with item 1, β_{2j} is the effect of item 2, and so on, β_{0j} is the intercept term. X_{qij} is the q^{th} dummy variable for person j with a value of 1 when q=i and 0 otherwise. The expected value and variance of item responses, Y_{ij}, are:

$$E(Y_{ij}/p_{ij}) = p_{ij} \text{ and } var(Y_{ij}/p_{ij}) = p_{ij}(1-p_{ij}),$$

where p_{ij} is the probability that j gets i correct (Kamata, 2001, p. 82).

In order to have a point of reference, one of the items is dropped (usually the last item but not necessarily the last item). Therefore equation 1 becomes,

$$\eta_{ij} = \beta_{oj} + \beta_{1j} X_{1ij} + \ldots + \beta_{(k-1)j} X_{(k-1)j} \quad (2)$$

$$= \log (p_{ij} /(1-p_{ij})) \quad (3)$$

where β_{0j} is the expected item effect of the dropped item for person j. To find the effect of item i that is associated with the q^{th} dummy variable (X_{qij} = 1 when q = i and 0 otherwise), where β_{0j} is the intercept and β_{qj} is the specific effect of the q^{th} dummy variable, equation 2 is reduced to

$$\eta_{ij} = \beta_{oj} + \beta_{qj} \quad (4)$$

Combining equations 3 and 4, we get

$$\log (p_{ij} /(1-p_{ij})) = \eta_{ij} = \beta_{oj} + \beta_{qj} \quad (5)$$

Rearranging equation 5, we get

$$p_{ij} = 1/(1 + \exp(-\eta_{ij})) \quad (6)$$

In the level-1 model which is a item level model, the βs are the item difficulties or item effects which are not constant across persons. However in the level-2 model, the βs are constant across persons whereas β_{oj} is assumed to be the random effect across persons (Kamata, 2001). According to the Rasch model developed using HLM, one software package for estimating multilevel models written by Kamata, the level-2 models in a HLM Rasch model are

$$\beta_{oj} = \gamma_{00} + u_{0j} \quad (7.a)$$

$$\beta_{1j} = \gamma_{10} \quad (7.b)$$

.

.

.

$$\beta_{(k-1)j} = \gamma_{(k-1)0} \quad (7.k)$$

where u_{oj}, the random component of β_{oj}, is normally distributed with a mean of 0 and variance τ and denotes the ability, attitude, or latent trait of the person j. The absence of the random component terms in equations 7.b through 7.k shows that the item parameters are fixed across persons and vary across items. As can be seen from equation 7.a, u_{oj}, which is the person ability, attitude, or latent trait, is fixed across items and varies across persons. Combining equations 4 and 7.a, we get (for i=q)

$$\eta_{ij} = \gamma_{00} + u_{0j} + \gamma_{q0} \quad (8)$$

To find the probability of person i getting item j correct, we combine equations 6 and 8 to get (when i=q)

$$p_{ij} = 1/[1 + \exp\{-(u_{0j} + \gamma_{00} + \gamma_{q0})\}] \quad (9)$$

The equation for Rasch model is

$$p_{ij} = 1/[1 + \exp\{-(\theta_j - \delta_i)\}] \quad (10)$$

Comparing 9 and 10 we can conclude that equation 9 is an equivalent of the Rasch model if $u_{0j} = \theta_j$ and $-\gamma_{00} - \gamma_{q0} = \delta_i$.

To take this up another model level, a third variable can be considered. Consider that the students belong to n different schools and the researcher wishes to examine the effect of the schools on student ability, attitude, or latent trait and item difficulty. This could be considered as another level of hierarchy. If the school the student belongs to is indicated by m, m would be added to the earlier equations and variables. η_{ij} would become η_{ijm}, p_{ij} would become p_{ijm}, β_{oj} would become β_{ojm}, γ_{00} would become γ_{00m} and u_{00} would become u_{00m}. In the third-level or the school-level model, only the overall effect of items, γ_{00m} would vary across schools. For school m, the model would be

$$\gamma_{00m} = \pi_{000} + r_{oom} \quad (11.a)$$

$$\gamma_{10m} = \pi_{100} \quad (11.b)$$

.

.

.

$$\gamma_{(k-1)0m} = \pi_{(k-1)00} \quad (11.k)$$

where π_{000} is the fixed component of γ_{00m} and r_{oom} is the random component of γ_{00m} with a mean of 0 and variance τ_{π}. Combining equations 9 and 11.a through 11.k, we get

$$p_{ijm} = 1/[1 + \exp\{-(r_{oom} + u_{0jm}) - (-\pi_{q00} - \pi_{000})\}] \quad (12)$$

where $-\pi_{q00} - \pi_{000}$ is the item difficulty for the item i=q, and π_{000} is the item difficulty for the reference item k. The person ability, attitude, or latent trait of j belonging to school m is $r_{oom} + u_{0jm}$ which can be divided into two parts, the random effect associated with school m (r_{oom}) and the average ability, attitude, or latent trait of students in school m (u_{0jm}). Thus the individual student's ability, attitude, or latent trait can also be compared to the average ability, attitude, or latent trait of the students in the school m (Kamata, 2001). IRT models which have hierarchically ordered variables are called Multilevel IRT (MLIRT) models.

Advantages of MLIRT Models

Why would a researcher prefer MLIRT to other IRT estimation techniques, given that the estimates yielded by MLIRT are comparable to the other IRT estimation techniques for both dichotomous (Kamata, 1998) and polytomous items (Williams, 2003)? In many IRT techniques, the item and person parameters are estimated simultaneously. This gives rise to the "Neyman-Scott problem," which is the inconsistency in the maximum likelihood estimates of item parameters when they are estimated simultaneously (Neyman & Scott, 1948). This happens because the number of person abilities or attitudes increases with increase in sample size (number of respondents – with each person having a certain attitude level). Therefore, when the sample size increases, the estimates of item parameters become inconsistent due to

insufficient statistics that are available for the person attitude/ability values. One of the advantages of using MLIRT is the ability to treat item parameters as fixed and person abilities as random parameters, thereby avoiding the Neyman-Scott problem (Kamata, 2001).

Multilevel formulation of IRT facilitates the modeling of multiple-group IRT models (Bock & Zimowski, 1997) and thereby its special cases, such as group-level IRT model (Mislevy, 1983; Mislevy & Bock, 1989), the item parameter drift model (Bock, Muraki, & Pfeiffenberger, 1988), and the duplex design model (Bock & Mislevy, 1989). Therefore the effects of the variables such as person or group-characteristics can be evaluated. In the two-level analysis, when person characteristics are taken into account, such as gender in the earlier example, the effect of those characteristics can be estimated. Similarly, the three-level analysis, when group membership and the hierarchical structure of the data are taken into account, estimates the effects of group-level and person-level abilities, the interaction effects of person characteristics and group membership, and the estimate of person-level effects across groups (Kamata, 2001; Williams, 2003). This provides additional information about the parameter estimates at each level of the model, thereby avoiding the need to perform separate analyses (Adams, Wilson, & Wu, 1997; Kamata, 1998).

Polytomous IRT Models

Although the research in IRT has come a long way over the decades, very little work has been done using data with more than two categories using polytomous IRT models (Adams, Wilson, & Wu, 1997; Maier, 2001; Rijmen, Tuerlinckx, & De Boeck,

2002). If a person is asked whether he/she is happy, it is almost always impossible to expect a normal person to be happy always or never. When such a dichotomous item is administered to a respondent, he/she is forced to choose a yes/no option whereas traits exist at all levels in between. There are different levels of happiness and, when given a choice of responses ranging from always, often, sometimes, and never, a more accurate measure of the happiness of the respondents would be obtained. As Ostini and Nering (2006) observed, "…polytomous items measure across a wider range of the latent trait continuum than do dichotomous items… The advantage of polytomous items is that, by virtue of their greater number of response categories, they are able to provide more information over a wider range of trait continuum" (pp. 7-8).

Dichotomous items are appropriate for right-wrong measures of knowledge. But Kamakura and Balasubramanian (1989) reported that, especially in the measurement of social and personality variables, dichotomous measurements are less clear and in order to understand the data clearly, "more subtle nuances of agreement/disagreement" are needed. Cox (1980) said that items with dichotomous response alternatives are inadequate because they do not transmit the necessary information and are frustrating to the respondents. There are different types of polytomous IRT models, such as Nominal Response Model (NRM) (Bock, 1972), Graded Response Model (GRM) (Samejima, 1969), Generalized Partial Credit Model (GPCM) (Muraki, 1992; 1997), and Rating Scale Model (RSM) (Andersen, 1973; Andrich, 1978).

Estimation of Polytomous IRT Parameters

The item parameters of polytomous models are estimated by treating the polytomous items as concatenated dichotomous items (Ostini & Nering, 2006). However, when polytomous models are involved, a new parameter comes into picture, the threshold parameter. A variation of the difficulty or the location parameter is the threshold parameter. Thresholds are simply the boundaries that separate the categories of responses. In other words, thresholds indicate the probability of crossing over from one choice in the response to the immediate next choice (either higher or lower in trait). Graphically speaking, the polytomous Item Characteristic Curves' chart area will now contain *n* ICCs, *n* being the number of choices per item. Each curve will indicate the probability of the respondent choosing that particular option for the item, as shown in Figure 1. The points of intersection of these curves indicate the threshold parameter. Fewer boundaries are needed to separate the response categories; to be precise, the number of boundaries required will be *n-1*. Thus each category boundary is modeled separately as a dichotomous model and then combined to form the information for the entire item (Ostini & Nering, 2006).

An often used polytomous IRT model is Samejima's (1969, 1997) graded response model (GRM), a generalization of the 2PL model that permits estimation of multiple b_{ij} parameters per item (j from 1 to n-1) associated with n response categories. The formula for a GRM trace line is:

$$P(x_i = j|\theta) = \frac{1}{1+\exp[-a_i(\theta - b_{ij})]} - \frac{1}{1+\exp[-a_i(\theta - b_{ij+1})]} \qquad (13)$$

which states that the probability of responding in category *j* is the difference between a 2PL trace line for the probability of responding in category *j* or higher and a 2PL trace line for the probability of responding in category *j*+1 or higher.

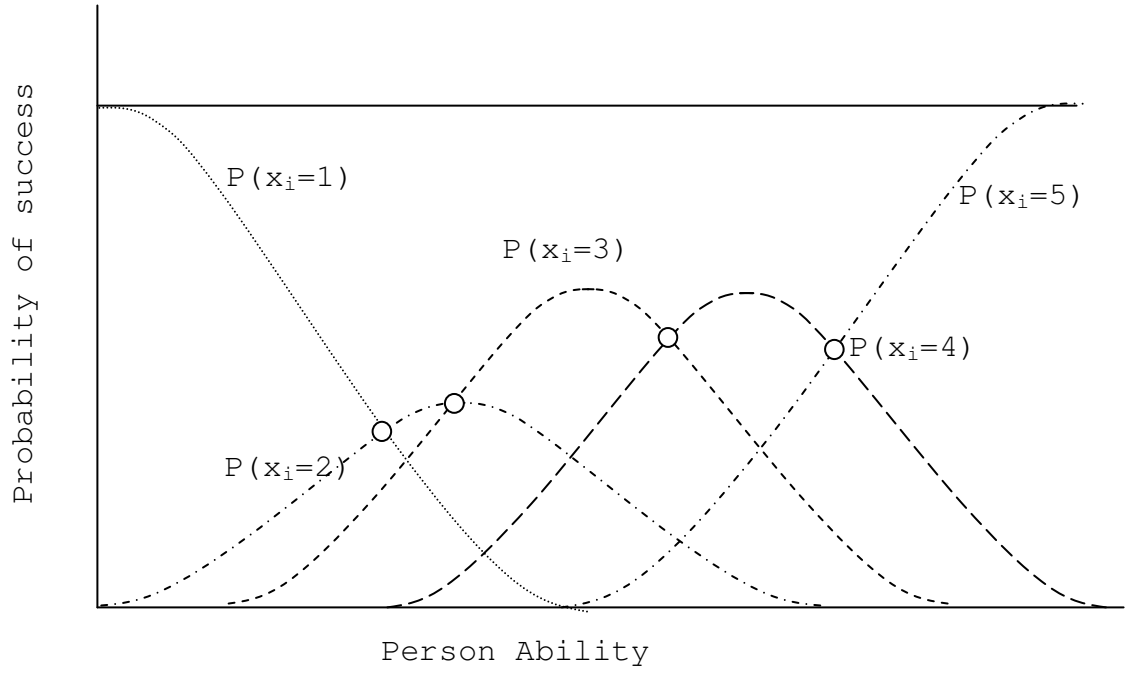


***Figure* 1. Information Function of a Polytomous Item with 5 Possible Response Options.**
The small circles at the intersection of the ICCs represent the 4 thresholds.

Monte Carlo Simulation for IRT Models

Simulation studies have been around even before the advent of high-speed computing when simulated data were formed based on the random drawing of a numbered ball or a piece of paper (Davey, Nering & Thompson, 1997). The most popular of the simulation studies, the Monte Carlo simulation, relates to or involves "the use of random sampling techniques and often the use of computer simulation to obtain approximate solutions to mathematical or physical problems especially in terms of a range of values each of which has a calculated probability of being the solution" (Merriam-Webster, Inc., 1994, pp. 754-755).

When a theory has to be validated by application to real data, it is useful to already know the estimates of the data, because these can then serve as the reference points to test for the accuracy of the estimated values. Typically, a researcher would want the estimated values to be as close as possible to the original distribution being simulated. Such a technique is especially useful when testing the theory for various data conditions and assumptions of the theory or their violations. This helps the researcher to understand the gravity of using the statistical theory for the situations encountered in real life datasets, such as violation of the homogeneity of variance assumption in ANOVA. One is able to estimate the performance of the theory by replicating the simulations numerous times in order to understand how statistical models behave in real life situations (Davey, Nering & Thompson, 1997; Fan, Felsövályi, Sivo, & Keenan, 2002).

Although simulation methods are very powerful in replicating situations and studying them, some situations simulations have been badly conducted (Hammersley &

Handscomb, 1964). This is due to the fact that, in spite of being an extremely powerful tool, caution must be exercised when simulating a dataset. Care has to be taken to generate the data as close as possible to real life data. For instance, a perfectly normal distribution is almost impossible to find in real life studies. Errors such as measurement error and sampling error, not to mention the model specification error, are common in real data sets. Because in real life many factors contribute to "noise" or lack of fit (especially in the behavioral sciences) and "simulations are useful only to the extent that they reflect reality" (Davey, Nering & Thompson, 1997, p. 4), lack of fit has to be introduced in the simulated data for increased validity.

While simulating a latent-trait based model, the form of the IRT model (1, 2, or 3-PL), the corresponding parameters, and the latent trait of the persons are usually necessary to adequately specify the simulated data. Item parameters could either be randomly drawn from probability distributions or selected from those estimated from items actually administered to people (Davey, Nering & Thompson, 1997). Some researchers also create parameters by fitting an MLIRT model to a large sample of real life data. This model then serves as a template to simulate other IRT models (e.g., Tucker, Koopman, & Linn, 1969). The present research utilized the Monte Carlo simulation to simulate data for the 2-PL MLIRT models. Unlike regular IRT simulation studies that simulate data and then estimate the parameters, the parameters of the simulated dataset here were predetermined to facilitate investigation of accuracy and precision of the simulated estimates.

Various researchers recommend different minimum test lengths and sample sizes when estimating parameters of an IRT model ranging from 30 to 80 for test lengths and 500 to 1000 for sample sizes (Hulin, Lissak, & Drasgow, 1982; Lord, 1968; Ree & Jensen, 1980; Swaminathan & Gifford, 1983). Therefore, test lengths of 15, 30, and 60 and sample sizes of 200, 500, 1000, and 2000 were studied because these sample sizes and test lengths extend above and below the minimum size recommended by Lord (1968) and Swaminathan and Gifford (1983).

Research has found that the three and four-choice items behave similarly with respect to item discrimination and consistency (Bruno & Dirkzwager, 1995; Rogers & Harley, 1999). However, difficulty (location) is inversely related to the number of categories and therefore four-choice items offer a wider range of options for testing and evaluating models. Instead of simulating numerous datasets for each condition, several iterations were run for each condition and the estimates from each iteration were used to calculate the accuracy and precision of the estimates. The typical values of discrimination and difficulty range from -2.8 to +2.8 and -3.00 to +3.00, respectively (Baker, 1985, 1992). However, Hambleton and Swaminathan (1991) stated that desirable items should not have negative discrimination values. Therefore, the simulated items were specified to have only positive discrimination values.

Methodology

Methodology for Simulation

Generation of responses was based on the discrimination and threshold parameters, the proportion of people in groups 1 and 2, the point biserial correlation between the responses and the group membership, sample size, test length, and the ability (or attitude level) parameter. A program was written in MATLAB 7.1, technical computing software which uses matrix algebra for all computations and modeling. This program was designed to generate the ability (or attitude level) of the person as a unit normal distribution.

The discrimination parameters for each item were generated based on the test length. The discrimination parameters of 0.75 and 1.33 correspond to the factor pattern coefficients of 0.6 and 0.8, which is necessary for a good test (Reise & Yu, 1990; Samejima, 1976). Therefore, in order to ensure unidimensionality, the slope values were generated with values equally spaced between 0.75 and 1.33. The first thresholds for the items were randomly chosen using a uniform random distribution with values between -3.0 and -1.0 (using the function *unifrnd*). Similarly, the second thresholds were randomly chosen between -1 and 1, and the third thresholds were randomly chosen between 1 and 3.0.

Before the responses were generated, the range of values that the mean and the variance of the abilities of the people belonging to groups 1 and 2 can take on were derived. The point biserial correlation, r_{pb} between a continuous variable x (ability/latent

trait, in this case) and a categorical variable with 2 groups 1 and 2 (predictor) is given by the formula:

$$r_{pb} = \frac{(\mu_1 - \mu_2)\sqrt{p(1-p)}}{\sigma_x} \tag{14}$$

where r_{pb} is the point biserial correlation, μ_1 and μ_2 are the population means of groups 1 and 2 respectively, p is the proportion of people in group1 to the total sample, and σ_x is the standard deviation of the variable x in the population.

Rewriting equation 14 for sample, we get:

$$r = \frac{(\overline{X}_1 - \overline{X}_2)}{s_x}\sqrt{\frac{n_2 n_1}{n(n-1)}} \tag{15}$$

where r is the point biserial correlation in the sample, n_1 and n_2 are the sample sizes of groups 1 and 2 respectively, n is the size of the sample with both groups 1 and 2 put together, and s_x is the standard deviation of the variable x in the sample. As groups 1 and 2 are the entire sample, $n_1+n_2=n$. Therefore equation 15 can be rewritten as,

$$r = \frac{(\overline{X}_1 - \overline{X}_2)}{s_x}\sqrt{\frac{n_2(n-n_2)}{n(n-1)}} \tag{16}$$

The distributions of the variables X_1 and X_2 are as follows:

$$X_1 \sim N(\mu_1, \sigma_1^2) \text{ and } X_2 \sim N(\mu_2, \sigma_2^2)$$

Therefore, the distribution of X can be given by,

$$X = (1-\omega)X_1 + \omega X_2$$

where ω represents group membership and takes on the value 1 when the datapoint belongs to group 2 or 0 when the datapoint belongs to group 1. Generalizing to any function f_x

$$f_x = (1\text{-}p)\, f_{x1} + p\, f_{x2} \tag{17}$$

Therefore, the mean and the variance of X are given by,

$$\mu_X = (1-p)\mu_1 + p\,\mu_2 \tag{18}$$

$$Var(X) = \int (X - \mu_1 + \mu_1)^2 (1-p) f_{x1} + \int (X - \mu_2 + \mu_2)^2 p f_{x2} \tag{19}$$

Integrating equation 19 over the limits of the variable X,

$$Var(X) = (1-p)Var(X_1) + \mu_1^2(1-p) + pVar(X_2) + \mu_2^2 p \tag{20}$$

$$E(X^2) = (1-p)\sigma_1^2 + p\sigma_2^2 + \mu_1^2(1-p) + \mu_2^2 p \tag{21}$$

$$E((X-\mu)^2) = (1-p)\sigma_1^2 + p\sigma_2^2 + (\mu_1 - \mu)^2(1-p) + (\mu_2 - \mu)^2 p \tag{22}$$

But $E((X-\mu)^2) = \sigma_X^2$. Additionally, in this case, the latent trait has a unit normal distribution which means $\sigma_X^2 = 1; \mu_X = 0$. Therefore, equation 18 reduces to,

$$\mu_X = (1-p)\mu_1 + p\,\mu_2 = 0$$

$$\Rightarrow (1-p)\mu_1 = -p\mu_2 \tag{23}$$

$$\Rightarrow \mu_1 = \frac{-p\mu_2}{(1-p)} \tag{24}$$

Substituting for the total variance equals 1, equation 22 reduces to,

$$1 = (1-p)\sigma_1^2 + p\sigma_2^2 + \mu_1{}^2(1-p) + \mu_2{}^2 p \tag{25}$$

$$1 = (1-p)\sigma_1^2 + p\sigma_2^2 + \frac{p^2}{(1-p)^2}(1-p)\mu_2{}^2 + \mu_2{}^2 p \tag{26}$$

$$1 = (1-p)\sigma_1^2 + p\sigma_2^2 + \frac{p^2}{(1-p)}\mu_2{}^2 + \frac{\mu_2{}^2 p(1-p)}{(1-p)} \tag{27}$$

$$1 = (1-p)\sigma_1^2 + p\sigma_2^2 + \frac{p}{(1-p)}\mu_2{}^2 \tag{28}$$

Rearranging equation 24, we get

$$\mu_1 - \mu_2 = (\frac{-p}{(1-p)} - 1)\mu_2 \tag{29}$$

$$\mu_2 - \mu_1 = \frac{\mu_2}{(1-p)} \tag{30}$$

Substituting for μ_2- μ_1 from equation 30 in equation 15, we get,

$$r = \frac{\mu_2\sqrt{p(1-p)}}{(1-p)} \tag{31}$$

$$\Rightarrow r = \mu_2\sqrt{\frac{p}{(1-p)}} \tag{32}$$

$$\Rightarrow \mu_2 = r\sqrt{\frac{(1-p)}{p}} \tag{33}$$

Substituting equation 33 in equation 28, we get,

$$1 = (1-p)\sigma_1^2 + p\sigma_2^2 + \frac{p}{(1-p)}r^2\frac{(1-p)}{p} \tag{34}$$

$$\Rightarrow 1 = (1-p)\sigma_1^2 + p\sigma_2^2 + r^2 \tag{35}$$

$$\Rightarrow 1 - r^2 = (1-p)\sigma_1^2 + p\sigma_2^2 \tag{36}$$

$$\Rightarrow \sigma_1^2 = \frac{((1-r^2)) - p\sigma_2^2}{(1-p)} \tag{37}$$

But the variance of X_1 must lie between 0 and 1. Therefore,

$$\frac{(1-r^2)-p\sigma_2^2}{(1-p)}>0 \tag{38}$$

$$\Rightarrow (1-r^2)-p\sigma_2^2>0 \tag{39}$$

$$\Rightarrow \sigma_2^2<\frac{1-r^2}{p} \tag{40}$$

and

$$\frac{(1-r^2)-p\sigma_2^2}{(1-p)}<1 \tag{41}$$

$$\Rightarrow 1-r^2-p\sigma_2^2<(1-p) \tag{42}$$

$$\Rightarrow -r^2-p\sigma_2^2<-p \tag{43}$$

$$\Rightarrow r^2+p\sigma_2^2>p \tag{44}$$

$$\Rightarrow \sigma_2^2>(1-\frac{r^2}{p}) \tag{45}$$

In sum, the variance was specified for the second group based on the point biserial correlation and the proportion of people in group 1 compared to the total sample. Therefore, the value for σ_2^2 was generated using the following constraint:

$$(1-\frac{r^2}{p})<\sigma_2^2<1-\frac{r^2}{p} \tag{46}$$

$$\max(0,(1-\frac{r^2}{p}))<\sigma_2^2<\min(1-\frac{r^2}{p},1) \tag{47}$$

The variance for group 1 was calculated based on the variance of group 2 using equation 34. The mean values for groups 1 and 2 were estimated using equations 24 and 32 respectively. Therefore, the predictor variable and the ability (or attitude level) values for the sample were generated using the statistics of groups 1 and 2 estimated in the above given derivation. The algorithms for simulation are given in Appendix A. The MLIRT model for the simulation study is graphically represented in Figure 2.

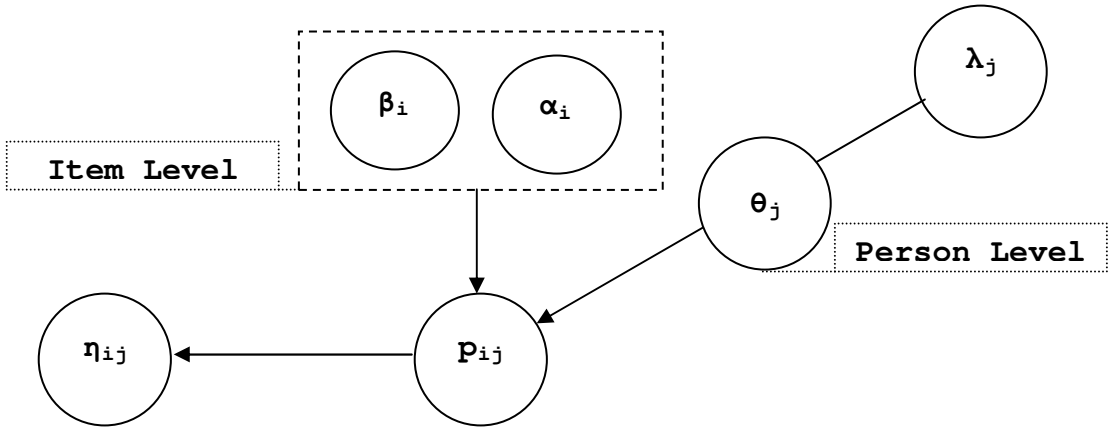


***Figure* 2. Graphical Representation of a 2-PL MLIRT Model with a Predictor Variable**

θ_j is the latent trait of person j; λ_j is the covariate of the latent trait of person j; β_i is the slope (discrimination parameter) of item i; α_i is the threshold parameter of item i; p_{ij} is the probability of person j answering item i; η_{ij} is the response given by person j for item i

Parameter Estimation

In the WINBUGS code, each parameter was estimated based on a given set of prior values. The slope was estimated by setting the distribution to a uniform distribution with values ranging between 0.5 and 1.5. Although the slope values for a unidimensional model should remain between 0.75 and 1.33, values outside this ideal range were tested to explore robustness of the models. The first threshold values were set to vary between -3 and 0, the second thresholds were allowed to vary between -1.5 and 1.5, and the third thresholds were allowed to vary between 0 and 3.0. The WINBUGS algorithm for estimation of parameters is given in Appendix B. The estimates from 1000 iterations were used for the analysis instead of estimates from 1000 different samples generated for the given condition.

Two Goodness of Recovery (GOR) measures proposed by Maris (1999), Bias and Root Mean Square Deviation (*RMSD*) were estimated. *Bias* is the difference between the average of the estimated parameter values and the true values. It is given by the formula,

$$BIAS(\tau_j) = (\Sigma\, \tau'_{jr}/R) - \tau_j$$

where τ_j is the true value of the parameter j, τ'_{jr} is the estimated value of the parameter for the r^{th} simulated dataset, and $r = 1,2,..R$. *RMSD* is the square root of the average squared differences between the estimated and the true parameter values. It is given by the formula,

$$RMSD(\tau_j) = SQRT(\Sigma\, (\tau'_{jr} - \tau_j)^2/R)$$

where τ_j is the true value of the parameter j, τ'_{jr} is the estimated value of the parameter, R is the number of replications. *RMSD* is a function of the Monte Carlo Standard Error (MCSE) of the simulated data and bias, given by the formula:

$$RMSD^2 = MCSE^2 + BIAS^2$$

Bias represented the accuracy of the estimates and RMSD represented the precision. In the best case scenario, the value of bias will be negligible and the distribution of RMSD will be identical to that of the simulated data set. Lower values of bias indicate better estimates. If 95% confidence intervals for the response probabilities (2 X 1.96 X RMSD) has an average width of less than 20% of the range of probability, statistically sufficient RMSD values are said to have been obtained (Maris, 1999). The Pearson correlation between the estimated and the true parameter values further indicated the extent and direction of bias. Therefore, the condition (minimal test length, sample size, and predictor distribution) at which optimal values of RMSD, bias, and correlation are obtained were identified. Additionally, factorial ANOVAs were conducted to quantify the effect of test lengths, sample sizes, binomial distribution of predictor variable, and the point biserial correlation between the predictor variable and the ability (or attitude level) of the individual on the estimates of the parameters. Because these tests were conducted to simply quantify the effects and not used as a tool to test statistical significance, inflation of the experimentwise Type I error rate was not a concern in this analysis.

Results

Test Length

Table 2 shows the results of the estimation from 1000 iterations for all the simulation conditions. From the results, it can be seen that the bias values for the discrimination parameter increase with increase in test length. The magnitude of the bias and RMSD values increase with increase in test length. Similarly, the Pearson's correlation (r) between estimated and simulated values for the discrimination parameter decreased with increase in test length which could be attributed to the higher number of parameters to be estimated with increased test length. The MCSE values did not improve with change in test length. Bias in threshold generally increased with increase in test length, which could also be attributed to the higher number of parameters to be estimated. There was almost no effect on the RMSD and the MCSE values of the threshold parameter.

Sample Size

Sample size had a weak effect on the bias of discrimination parameters. The bias values of the discrimination parameters decreased with increase in sample size but no strong pattern was detected. However, the RMSD and MCSE values of the discrimination parameter decreased with increase in sample size, which can be expected because RMSD and MCSE are analogous to standard deviation and standard error respectively. Higher sample sizes yielded better estimates which tend to fluctuate less between iterations. The correlation between estimated and simulated discrimination values increased with increase in sample size. This could be attributed to better and less fluctuating estimates at higher sample sizes.

Distribution of the Predictor Variable

In general, the distribution of the binomial predictor variable did not have an impact on the estimates of both discrimination and threshold parameters. Irrespective of the ratio between group sizes (0.1:0.9, 0.25:0.75, 0.4:0.6, and 0.5:0.5) there was no detectable pattern between the estimates of the discrimination and threshold parameters and the binomial distribution.

Table 2
BIAS, RMSD, and Pearson's Correlation (r) for Discrimination and Threshold Parameters for Simulated Conditions

Test Length	Sample Size	p	r_{pb}	Discrimination				Threshold			
				BIAS	RMSD	MCSE	r	BIAS	RMSD	MCSE	r
15	200	0.10	0.30	−0.054	0.177	0.168	0.834	−0.049	0.762	0.760	0.861
		0.10	0.85	−0.018	0.187	0.186	0.706	0.000	0.744	0.744	0.859
		0.25	0.30	0.000	0.176	0.176	0.834	−0.073	0.765	0.761	0.860
		0.25	0.85	−0.018	0.169	0.168	0.798	−0.004	0.772	0.772	0.855
		0.40	0.30	0.010	0.176	0.178	0.865	−0.005	0.772	0.772	0.850
		0.40	0.85	−0.006	0.165	0.165	0.872	−0.011	0.782	0.782	0.853
		0.50	0.30	−0.054	0.149	0.139	0.951	−0.054	0.823	0.821	0.847
		0.50	0.85	0.015	0.184	0.183	0.738	−0.004	0.723	0.723	0.863
	500	0.10	0.30	−0.022	0.167	0.165	0.863	−0.134	0.663	0.649	0.908
		0.10	0.85	−0.013	0.112	0.111	0.941	−0.115	0.623	0.612	0.917
		0.25	0.30	−0.027	0.120	0.117	0.910	−0.136	0.652	0.637	0.910
		0.25	0.85	−0.010	0.118	0.117	0.918	−0.116	0.595	0.583	0.919
		0.40	0.30	0.030	0.110	0.105	0.952	−0.122	0.605	0.592	0.920
		0.40	0.85	0.020	0.119	0.117	0.892	−0.151	0.635	0.617	0.909
		0.50	0.30	−0.023	0.117	0.115	0.930	−0.118	0.627	0.616	0.923
		0.50	0.85	−0.032	0.127	0.123	0.846	−0.108	0.644	0.635	0.918
	1000	0.10	0.30	0.025	0.089	0.085	0.941	−0.016	0.147	0.146	0.997
		0.10	0.85	0.017	0.094	0.092	0.939	0.013	0.153	0.152	0.997
		0.25	0.30	0.028	0.100	0.096	0.933	0.005	0.145	0.145	0.997
		0.25	0.85	0.015	0.089	0.088	0.945	−0.039	0.157	0.152	0.997

Table 2 (continued)

Test Length	Sample Size	p	r_{pb}	Discrimination BIAS	Discrimination RMSD	Discrimination MCSE	Discrimination r	Threshold BIAS	Threshold RMSD	Threshold MCSE	Threshold r
15	1000	0.40	0.30	−0.033	0.095	0.089	0.948	0.022	0.174	0.173	0.996
		0.40	0.85	−0.008	0.098	0.098	0.941	0.009	0.176	0.176	0.996
		0.50	0.30	0.026	0.101	0.097	0.925	0.013	0.138	0.137	0.998
		0.50	0.85	0.000	0.095	0.095	0.931	−0.003	0.175	0.175	0.995
	2000	0.10	0.30	0.005	0.060	0.060	0.984	−0.107	0.737	0.729	0.866
		0.10	0.85	0.006	0.069	0.069	0.959	−0.097	0.735	0.728	0.865
		0.25	0.30	0.008	0.071	0.070	0.964	−0.124	0.760	0.750	0.858
		0.25	0.85	−0.030	0.078	0.072	0.955	−0.110	0.753	0.745	0.864
		0.40	0.30	−0.024	0.077	0.073	0.946	−0.128	0.745	0.734	0.867
		0.40	0.85	−0.022	0.079	0.076	0.939	−0.125	0.748	0.737	0.863
		0.50	0.30	0.009	0.066	0.066	0.968	−0.101	0.722	0.715	0.868
		0.50	0.85	0.015	0.073	0.071	0.950	−0.126	0.723	0.712	0.866
30	200	0.10	0.30	−0.112	0.195	0.160	0.765	−0.004	0.363	0.363	0.985
		0.10	0.85	−0.114	0.191	0.153	0.809	0.038	0.381	0.379	0.985
		0.25	0.30	−0.129	0.197	0.149	0.754	−0.012	0.345	0.345	0.988
		0.25	0.85	−0.102	0.175	0.142	0.772	0.014	0.339	0.339	0.988
		0.40	0.30	−0.103	0.179	0.146	0.798	0.043	0.336	0.333	0.989
		0.40	0.85	−0.112	0.185	0.147	0.763	0.044	0.340	0.337	0.987
		0.50	0.30	−0.085	0.185	0.164	0.709	0.040	0.345	0.342	0.987
		0.50	0.85	−0.083	0.188	0.169	0.725	0.030	0.351	0.350	0.985

Table 2 (continued)

Test Length	Sample Size	p	r_{pb}	Discrimination BIAS	Discrimination RMSD	Discrimination MCSE	Discrimination r	Threshold BIAS	Threshold RMSD	Threshold MCSE	Threshold r
30	500	0.10	0.30	−0.098	0.144	0.105	0.848	−0.095	0.694	0.687	0.905
		0.10	0.85	−0.042	0.130	0.123	0.873	−0.075	0.658	0.654	0.910
		0.25	0.30	−0.067	0.128	0.109	0.839	−0.086	0.707	0.702	0.901
		0.25	0.85	−0.039	0.141	0.135	0.813	−0.088	0.683	0.677	0.903
		0.40	0.30	−0.072	0.151	0.133	0.770	−0.083	0.662	0.657	0.906
		0.40	0.85	−0.080	0.146	0.122	0.799	−0.098	0.714	0.707	0.901
		0.50	0.30	−0.090	0.149	0.119	0.821	−0.087	0.693	0.687	0.905
		0.50	0.85	−0.057	0.134	0.121	0.842	−0.085	0.664	0.658	0.907
	1000	0.10	0.30	−0.068	0.089	0.057	0.860	−0.029	0.699	0.698	0.898
		0.10	0.85	−0.057	0.124	0.110	0.783	−0.007	0.687	0.690	0.899
		0.25	0.30	−0.044	0.106	0.096	0.855	−0.009	0.687	0.690	0.896
		0.25	0.85	−0.032	0.110	0.105	0.826	−0.002	0.664	0.664	0.898
		0.40	0.30	−0.048	0.116	0.106	0.803	0.000	0.693	0.693	0.894
		0.40	0.85	−0.070	0.122	0.100	0.833	−0.015	0.697	0.697	0.896
		0.50	0.30	−0.052	0.117	0.105	0.815	−0.016	0.684	0.684	0.900
		0.50	0.85	−0.067	0.114	0.092	0.845	0.005	0.704	0.704	0.897
	2000	0.10	0.30	−0.050	0.087	0.071	0.912	0.016	0.632	0.632	0.898
		0.10	0.85	−0.067	0.091	0.061	0.902	0.022	0.642	0.642	0.902
		0.25	0.30	−0.048	0.086	0.071	0.911	0.017	0.634	0.634	0.902
		0.25	0.85	−0.057	0.090	0.070	0.912	0.020	0.640	0.640	0.897
		0.40	0.30	−0.049	0.091	0.077	0.889	0.032	0.624	0.623	0.900
		0.40	0.85	−0.056	0.091	0.072	0.898	0.037	0.638	0.637	0.900

Table 2 (continued)

Test Length	Sample Size	p	r_{pb}	Discrimination BIAS	Discrimination RMSD	Discrimination MCSE	Discrimination r	Threshold BIAS	Threshold RMSD	Threshold MCSE	Threshold r
30	2000	0.50	0.30	−0.048	0.085	0.070	0.912	0.032	0.629	0.628	0.900
		0.50	0.85	−0.057	0.098	0.080	0.882	0.028	0.639	0.638	0.898
60	200	0.10	0.30	−0.093	0.194	0.170	0.634	−0.007	0.351	0.351	0.986
		0.10	0.85	−0.077	0.178	0.160	0.744	−0.017	0.334	0.333	0.988
		0.25	0.30	−0.089	0.193	0.171	0.647	−0.007	0.341	0.341	0.987
		0.25	0.85	−0.089	0.179	0.155	0.752	−0.014	0.328	0.328	0.989
		0.40	0.30	−0.082	0.181	0.161	0.691	0.003	0.330	0.330	0.986
		0.40	0.85	−0.080	0.179	0.160	0.756	0.006	0.323	0.323	0.989
		0.50	0.30	−0.078	0.178	0.194	0.628	−0.019	0.332	0.331	0.987
		0.50	0.85	−0.102	0.202	0.174	0.637	0.006	0.355	0.355	0.987
	500	0.10	0.30	−0.080	0.151	0.128	0.758	0.000	0.672	0.672	0.894
		0.10	0.85	−0.096	0.150	0.115	0.819	−0.005	0.657	0.657	0.901
		0.25	0.30	−0.100	0.152	0.114	0.819	−0.005	0.669	0.669	0.899
		0.25	0.85	−0.078	0.148	0.126	0.773	−0.014	0.652	0.652	0.900
		0.40	0.30	−0.071	0.147	0.129	0.783	0.017	0.653	0.653	0.899
		0.40	0.85	−0.070	0.134	0.114	0.809	−0.005	0.661	0.661	0.898
		0.50	0.30	−0.073	0.141	0.121	0.807	−0.025	0.656	0.655	0.899
		0.50	0.85	−0.087	0.152	0.125	0.778	−0.022	0.664	0.664	0.900
	1000	0.10	0.30	−0.051	0.112	0.100	0.863	−0.020	0.611	0.611	0.908
		0.10	0.85	−0.066	0.118	0.098	0.845	−0.017	0.598	0.598	0.909
		0.25	0.30	−0.062	0.119	0.101	0.838	−0.016	0.589	0.589	0.913

Table 2 (continued)

Test Length	Sample Size	p	r_{pb}	Discrimination BIAS	Discrimination RMSD	Discrimination MCSE	Discrimination r	Threshold BIAS	Threshold RMSD	Threshold MCSE	Threshold r
60	1000	0.25	0.85	−0.039	0.113	0.106	0.823	−0.022	0.593	0.592	0.910
		0.40	0.30	−0.054	0.109	0.095	0.857	−0.003	0.599	0.599	0.910
		0.40	0.85	−0.056	0.120	0.106	0.851	−0.008	0.599	0.599	0.910
		0.50	0.30	−0.053	0.112	0.099	0.834	−0.010	0.599	0.599	0.906
		0.50	0.85	−0.048	0.111	0.100	0.860	−0.027	0.581	0.580	0.914
	2000	0.10	0.30	−0.058	0.098	0.079	0.867	−0.012	0.613	0.613	0.901
		0.10	0.85	−0.062	0.101	0.080	0.852	−0.006	0.625	0.625	0.900
		0.25	0.30	−0.062	0.095	0.072	0.871	−0.001	0.615	0.615	0.901
		0.25	0.85	−0.062	0.098	0.076	0.861	−0.022	0.620	0.620	0.900
		0.40	0.30	−0.061	0.094	0.071	0.870	−0.017	0.620	0.620	0.901
		0.40	0.85	−0.055	0.094	0.076	0.872	−0.015	0.613	0.613	0.902
		0.50	0.30	−0.065	0.099	0.075	0.867	−0.006	0.616	0.616	0.902
		0.50	0.85	−0.058	0.096	0.076	0.875	−0.017	0.616	0.616	0.900

Note: Bias is the measure of accuracy of the estimates of the parameters across iterations. RMSD is the measure of consistency of the estimates of the parameters across iterations.

Table 3

η^2 Values from Factorial ANOVAs for the Simulation Design Features Explaining the Variabilities in Discrimination and Threshold Parameters

Simulation Main & Interaction Effects	Discrimination			Threshold		
	BIAS	RMSD	r	BIAS	RMSD	r
Test Length (TL)	67.1	5.4	32.2	26.3	0.7	2.5
Sample Size (SS)	13.3	89.6	47.9	30.4	24.5	24.0
Distribution (p)	0.2	0.1	0.4	0.6	0.0	0.0
Correlation (r)	0.1	0.0	0.0	0.1	0.0	0.0
TL X SS	3.2	0.2	3.3	35.7	74.1	73.1
TL X p	0.4	0.3	1.0	0.4	0.0	0.0
TL X r	0.0	0.0	1.5	0.3	0.0	0.0
SS X p	2.9	0.4	1.7	0.9	0.0	0.0
SS X r	0.6	0.3	0.4	0.8	0.0	0.0
p X r	0.2	0.1	0.4	0.7	0.0	0.0
TL X SS X p	5.3	0.9	2.8	1.1	0.1	0.1
TL X SS X r	1.1	0.3	3.1	0.5	0.0	0.0
TL X p X r	1.2	0.2	0.8	0.3	0.0	0.0
SS X p X r	1.2	1.1	2.2	0.6	0.1	0.1
TL X SS X p X r	0.0	0.0	0.0	0.0	0.0	0.0

Note: The η^2 values were calculated by dividing the sum of squares of the corresponding effect by the total corrected sum of squares (hence the effect sizes for each factorial ANOVA add up to 100%).

Point Biserial Correlation

The point biserial correlation between the ability (or attitude level) and the predictor variable did not have an impact on the estimates of either discrimination or threshold parameters. For both higher and lower correlation (0.3 and 0.85) there was no detectable pattern between the estimates of the discrimination and threshold parameters and the binomial distribution.

Overall Effects

Six factorial ANOVAs were conducted to investigate the main and interaction effects of the simulation conditions (3 Test Lengths X 4 Sample Sizes X 4 Binomial Distributions X 2 Point Biserial Correlations) on the estimates of the bias, RMSD, and the Pearson's correlation between the estimated and simulated values of discrimination and threshold parameters individually. The results of the factorial ANOVAs as shown in Table 3, indicate that while test length had the maximum effect on the bias of the discrimination parameter (67.1%), it had a very low effect on the RMSD (5.4%) but a moderate effect on the Pearson's correlation values of the discrimination parameter (32.2%). Sample size had the highest effect on the RMSD and Pearson's r (89.6% and 47.9% respectively) and also explained about 13.3% of the variance in the bias of the discrimination parameter.

Except the 2-way interaction effect between test length and sample size, all 2-way interaction effects explained less than 3% of the variance in the estimates of the discrimination and the threshold parameters. Most of the variation in the estimates of the threshold parameter was explained by the 2-way interaction between test length and

sample size conditions. The 2-way interaction between test length and sample size had an effect size of 35.7% for the bias, 74.1% of the RMSD, and 73.1% of the Pearson's correlation between the estimated and the simulated values of the threshold parameter. All 3-way interaction effects explained less than 5% of the variance in the estimates of the discrimination and the threshold parameters and the 4-way interaction effect had no impact on the estimates. While test length explained 26.3% of the bias, 0.7% of the RMSD, and 2.5% of the Pearson's correlation, sample size explained 30.4% of the bias, 24.5% of the RMSD, and 24% of the Pearson's correlation between the simulated and the estimated values of the threshold parameter.

Conclusion

The most important finding of this article is the least restriction 2-PL MLIRT models place on the predictor variable. The binomial distribution of the predictor variable and the point biserial correlation between the ability (or attitude level) of the individual and the predictor variable had minimal effect on the estimates of discrimination and threshold parameters. The practical implication of this finding is *that group sizes need not be comparable to yield good estimates of threshold and discrimination parameters*. Minimal effect of point biserial correlation on the estimates of the parameters implies that *2-PL MLIRT analysis can be done irrespective of relationship between the latent trait of the individuals to group membership.*

As expected, test length and sample size had the highest impact on the estimation of the threshold and the discrimination parameters. In general, the discrimination parameters were underestimated, especially when the test lengths were longer (30 and

60). However, when the sample sizes were high enough (1000 and 2000), the bias for longer tests were minimized. The RMSD and Pearson's r values for discrimination parameters increased when longer tests were combined with small sample sizes. The bias values of the threshold parameters were desirably low except when the test length was 15.

Some recommendations for robust test conditions for 2-PL MLIRT models are given here based on the estimates from the WINBUGS program and the factorial ANOVAs. The estimates of discrimination and threshold parameters were adequate for almost all test lengths (15, 30, and 60) and sample sizes (200, 500, 1000, and 2000). Although the threshold parameters fluctuated wildly (high RMSD values) when sample sizes were low (200) or when the test lengths were short (15), better estimates were obtained when more than 300 iterations were performed. This shows that, when several iterations (usually over 300) are run, adequate estimates are obtained. The bias, RMSD, and Pearson's r values of discrimination parameters remained adequate, for all sample sizes and test lengths.

The general recommendations for test lengths of 30 to 80 items and minimum required sample sizes of 500 to 1000 (Hulin, Lissak, & Drasgow, 1982; Lord, 1968; Ree & Jensen, 1980; Swaminathan & Gifford, 1983), based on the finding of this study seem reasonable. Here, test lengths of 15 items and sample sizes of 200 yield good estimates of discrimination and threshold parameters given adequate iterations. These conditions have lesser restrictions compared to the recommended test lengths and sample sizes because the presence of an additional predictor contributes more information about the ability (or attitude level) of the individual. Additionally, 2-PL MLIRT analysis puts minimum restrictions on the conditions for distribution of the predictor variable. Therefore, 2-PL MLIRT models offer a leap over the usual IRT models by providing *more power to yield better estimates at smaller test lengths and sample sizes.*

CHAPTER IV

A MULTILEVEL ITEM RESPONSE THEORY ANALYSIS OF URBAN TEACHERS' PERCEPTIONS OF THEIR CULTURAL AWARENESS AND BELIEFS IN TEACHING AFRICAN AMERICAN STUDENTS

Introduction

Students of color will make up about 46% of the nation's student population by the year 2020 (Banks, 1997). The National Center for Educational Statistics (NCES) also reported such an increase in the nation's student population (NCES, 2006). Although the school age population is becoming more diversified, the teachers of these students are predominately European-American, middle-class, and female (Miller, Miller, & Schroth, 1997; Taylor, 2000). Taylor (2000) reported that even though students of color are increasing, only 8% of public school teachers are African American and 6% of public school teachers are Hispanic (Digest of Education Statistics, 2003). Even these teachers come from an economic background different from the students they teach. Within their careers teachers will most likely have students from diverse ethnic, cultural, and racial groups in their classrooms (Banks, 1997). This leads to an economic-cultural mismatch between the schools and the home culture of the students (Garcia, 2001; Howard, 2001).

The impact is evident from the academic achievement gap between students of color and European American students. The reading level of an average 17-year old African American male remains equivalent to the reading ability of an average 13-year old White male (Anderson, 2004). African Americans make up 17% of the total student enrollment, yet account for 11% of the dropouts while Hispanic students make up 44%

of the total enrollment and account for 27% of the dropouts (Digest of Education Statistics, 2003). Although, historically the African Americans have overcome the various achievement gaps, such as the literary gap, the elementary gap, and the high school completion gap (Anderson, 2004), they are yet to overcome the test score gap. The achievement and educational attainment rates for African American students have improved in the recent years, but still remain below acceptable levels on National Standardized Tests (NCES, 2006). In all standardized tests, African Americans lag substantially behind their European counterparts (Bradford, Pitts, & Collins, 2002; Digest of Education Statistics, 2003; Green, 2001; Irvine & Armento, 2001; Jencks & Phillips, 1998).

In such a diverse environment, can a teacher teach every student, especially a student whose home culture and learning style is different from the mainstream Eurocentric culture and learning style, with equal effectiveness? This is one of the most challenging questions that education researchers and teacher educators face today. Over the past few decades, several researchers have shown the effectiveness of culturally responsive pedagogy in teaching students from diverse backgrounds. Such pedagogical methods, instruction methods, and culturally responsive curriculum helps increase the educational achievement of students of color (Banks, 1991a, 1993a, 1993b, 1994a; Banks & Banks, 2001; Boykin & Cunningham, 2001; Cohen, 1994; Delpit, 1995; Foster, 1992; Gay, 2002; Irvine, 1990; Ladson-Billings, 1990, 1994, 1995; Sleeter, 1995). In fact, Banks and Banks (2001) noted that continuing education about diversity is important for teachers especially because of the increasing ethnic and cultural gap

between the nation's teaching and the learning society. School success of African American students has been strongly linked to culture (Boykin, 1991).

Understanding the needs of diverse learners is one of the foremost challenges for teachers. According to many scholars (Gay, 2000; Howard, 2001; Ladson-Billings, 1994; Love, 2001; Villegas & Lucas, 2002) teachers' knowledge and implementation of culturally responsive pedagogy can impact and enhance the academic performance of students of color. However, for effective implementation of culturally responsive pedagogy, it is necessary to first know and understand the perceptions of teachers on cultural awareness and beliefs. Webb-Johnson and Carter (2005) developed an instrument that examined the cultural awareness and beliefs of urban teachers in order to develop intervention programs to help teachers with their pedagogical practices and to help narrow the gap between the learning styles of diverse learners and the teaching styles of teachers.

Most of the current researches have used classical test theory and therefore, although the real intent would be to measure the latent trait of the person (such as attitudes, perceptions, quality), researchers usually tend to restrict the analysis to hypothesis testing or measuring relationships among measured variables. In such cases, the hypotheses would simply compare if the sample under study performs better than another group. In short, the aim of measuring the trait is accomplished. Classical Test Theory (CTT) does not give an exact quantification of the trait itself. Item Response Theory (IRT), on the other hand, can give an exact measure of the trait of the person. When estimating a teacher's cultural awareness, item parameters would translate into the

contribution of each item to the awareness of the teacher. Items with higher IRT difficulties indicate lower endorsement rates by teachers. The impact of each item on the cultural awareness of a teacher is different, unlike in CTT, where each item is given equal weightage or importance. Thus, by giving each item a differential share in contributing to the cultural beliefs of teachers, an effective measure of latent traits can be obtained. Moreover, the relationship between the items can also be studied, by studying the item parameters. Therefore the questions that this study seeks to answer are:

1. What are the characteristics of the items that measure teachers' perceptions of cultural awareness and as shown by the IRT analysis using the 2-PL MLIRT model?
2. Do these item characteristics of cultural awareness differ by gender, ethnicity, age, and the years of teaching of the teachers?

Conceptually, the present study is useful for researchers in many fields and not simply researchers in testing because it demonstrates the use of IRT models for estimating latent traits. Although the model has been widely known and used in other disciplines, such as health, business, and marketing the use of IRT in mainstream educational research has been confined to mainly testing. Hopefully, the present study will increase the awareness of the applications of the IRT model in fields other than educational testing.

The Importance of Cultural Awareness: Induction and Professional Development

Gay (2000) defined culture as a multidimensional and continually changing concept which is influenced by time, setting, age, economics and social circumstances.

Culture influences a person's behaviors, such as thinking, relating, speaking, reading, writing, performing, producing, learning, and teaching styles. Shade, Kelly, and Oberg (1997) referred to the cultural characteristics most likely to be found in a group of people belonging to the same ethnic group as a *modal personality*. It is necessary for the teacher, the curriculum and the standards to be set to appreciate the differences among students of color, because such cultural differences guide a student's personality and learning to a great extent. The performance of students of color improves drastically when the instruction, teacher attitudes and expectations, and culturally relevant content come together to cater to the needs and the racial mix of the students.

Over the past decade, various exemplary teaching strategies have been developed for educating students of color (Delpit, 1995; Foster, 1992; Garcia, 2001; Howard, 2001; Irvine, 1990). However, this has not narrowed the achievement gap appreciably (Gay, 1995; Sleeter, 2001). A reason for this is because many teachers still do not understand the need to alter their pedagogical practices for students with different learning styles or students from different cultures. In fact, Phuntsog (2001) showed in a study of 66 elementary teachers that none wanted to incorporate multicultural education into the current curricula, content or the process of teacher education. Miller, Miller, and Schroth (1997) found that pre-service teachers lack knowledge about cultural issues and also sensitivity to the needs of diverse learners from different cultures. Additionally, they found that most pre-service teachers do not address issues related to gender, ethnicity, sexual orientation, social class, or cultural differences. Rothernberg, McDermott, and Gormley (1993) reported that many novice teachers do not understand the importance of

culturally responsive pedagogy or the interaction between culture and teaching. Therefore, there is an urgent need to develop strategies and induction programs to help teachers understand the importance of culture and its impact on learning style and also to develop strategies to help develop pedagogical practices and materials to suit the learning needs of diverse learners.

Several programs, such as the Teacher Induction Program (Moon-Merchant & Carter, 2004), have been developed throughout the country to help teachers manage their day-to-day problems. Many of these programs have been effective in reaching out to teachers who need assistance when they are new to the profession. However, when it comes to understanding the cultural mismatch and helping teachers develop pedagogical approaches to suit the needs of diverse learners, there still remains a pressing need for an effective induction program.

In order to develop an effective teacher induction program that can assist teachers with respect to multicultural understanding, first the existing beliefs of teachers must be understood. A notable observation is to be made at this juncture. Several qualitative studies have been conducted to understand and investigate the concerns of teachers in this area (Ladson-Billings, 1994). However, there is a need for quantitative studies because very few quantitative studies have been conducted with respect to the measurement of teacher beliefs and cultural awareness. Such a study would include several dimensions of attitudes of teachers towards their cultural awareness and beliefs, such as Teacher Beliefs about African American students, School Climate, Culturally Responsive Management, Home and Community Support, Cultural Sensitivity,

Curriculum and Instructional Strategies, Cultural Awareness, and Teacher Efficacy. The present study will discuss the analysis and findings pertaining to the teacher beliefs factor in detail as an example and also list the findings of the analyses of other factors. The *teacher beliefs about African American students* factor was discussed in detail due to various reasons, the most important reason is that this factor explained most of the variance (7.78%) as shown in Appendix C. It also contained the most number of items (8).

Teacher Beliefs about African American Students

A person's beliefs affect his/her ability to efficiently perform a specific task. Teachers' beliefs have been strongly linked to their behavior, perceptions, efficacy, and practices in the classroom by several researchers (Bandura, 1986; Brown, 2004; Rokeach, 1968). Reflective self-analysis is necessary for teachers who successfully implement equity pedagogy. This requires teachers to examine their attitudes towards people of different race, class, gender, and ethnic groups. In fact, King (1992) states that most teachers are unaware of the extent to which they embrace racist and sexist attitudes. Miller, Miller, and Schroth (1997) found that African American elementary students perceived teachers as having preferences for students who exhibit classroom behaviors consistent with competitive and individualistic mainstream cultural themes.

When teachers are made to believe that students from certain groups are difficult to teach, their attitudes towards teaching those students changes and decreases their sense of efficacy in educating these students effectively. This stems from what researchers call the deficit model, or the victim blame model. The proponents of this

model believe that some children, such as the children of color, are biologically or genetically inferior to other children and this is the reason behind their academic failure (Banks & Banks, 2001; Nieto, 2000, 2004). Some proponents of this model also believe that this is due to inadequate parenting and/or poverty (King, 2004; Pang & Sablan, 1998). Scholars who believe in the deficit model believe that it would be "extremely difficult" or "impossible" to close the achievement gap (Singham, 1998).

According to Banks (1988), a teacher's perception affects a student's performance. A study by Rosenthal and Jackson (1965) proved that teachers who believe some students to be higher achievers treat those students differently and these students had unusually high achievement scores. This shows that when the teacher believes a student can perform well, he/she gives more attention and encouragement to that student which causes high self efficacy for the student and therefore higher achievement scores. Therefore, such perceptions of teachers affect their efficacy or the extent the teachers believe they can actually teach children and make a difference in their lives (Bandura, 1997).

Many beginning teachers are trained in the deficit model and have a low sense of efficacy when it comes to teaching underserved students (Nieto, 2000). Irvine (1990, pp.7) reported that, "teachers form inaccurate impressions of student achievement especially with Black students." Teachers see students with African American culture-related or vervistic movements as academically less achieving, aggressive, and more likely to need special education services than other students (Neal, McGray, Webb-Johnson, & Bridget, 2003). Some teachers also believe that the circumstances in

students' lives prevent them from learning (Pang & Sablan, 1998). Lipman (1995) conducted a qualitative study of three successful teachers of African American students that showed that teachers had high academic and behavior standards for their students which helped them achieve these standards.

African American students bring with them certain strengths to the classroom that are almost never capitalized and built upon by teachers. Common examples include the use of "non-standard English" spoken by African American students that is rich, diverse, and commonly used in several social settings such as the work place and the community (Hale-Benson, 1986). In fact, Hollins and Spencer (1996) observed that most students of color are evaluated based on how close their behaviors are to the common middle class, White standards of achievement and behavior.

Ford and Grantham (1998) observed that such deficit models exist when teachers have negative and stereotypical views about students from diverse cultures which lower their expectations of their students. When teachers show lower expectations for students, the academic achievement and morale of students reduce significantly. In their study, Pohan and Aguilar (2001) found that pre-service teachers with negative bias towards students of color were less likely to develop professional beliefs and positive behaviors that embrace multicultural sensitivity and awareness. Irvine's (1990) study found that teachers formed impressions about their African American male students based on stereotypes rather than their achievement. Such attitudes prevent teachers from believing that students from diverse backgrounds can excel in learning and bring a variety of knowledge and skills to the classroom (Milner, 2005).

Several teachers also believe that African Americans with positive attitudes toward high achievement are perceived by their peers to be "acting White". According to this scholarship, African American students develop negative attitudes toward education, high achievement, and toward high academically achieving African American students (Fordham, 1988, 1999). According to this view, African Americans perceive high academic achievement as a mainstream value that benefits White Americans. Therefore, their academic achievement difficulties have been attributed to their attitudes towards high achievers (Ogbu, 1986). However, a study by Sankofa, Hurley, Allen, and Boykin (2005) showed that children rated the high achievers differently, depending on the cultural orientation of the high achievers. Black children were more accepting of their high achieving peers except when those peers' achievements reflected attitudes and behaviors incongruent with their own cultural orientations. They felt that their parents would perceive the same way as well. This study by Sankofa et al. (2005) contradicted the common belief that African American children reject academic achievement, as suggested by Ogbu (1986).

A better model that has the potential to narrow the achievement gap is the *cultural difference model* (Banks, 1988). The proponents of this model believe that the African Americans have a rich cultural heritage that is different from and not inferior to the European culture. The wide variety of cultural knowledge and experiences that students bring to the learning arena is a wealth that should be tapped upon. Moll and Gonzalez (2004) termed these skills as “funds of knowledge” (p. 702). In order to teach diverse learners effectively, a teacher has to understand the needs of diverse students,

their learning styles, and more importantly believe that all children bring with them a diverse wealth of knowledge.

Factors II and IV: School Climate and Home and Community Support

The term school climate refers to the psychological factors in a school context that affects student-teacher relationships (Kelly, Thornton, & Daughtery, 2005). School climate influences teacher, staff, and student behaviors (Hoy & Miskel, 2005), and describes issues that can affect staff attitudes and effectiveness (Esposito, 1999). School climate is an important determinant of how students form perceptions of themselves (Banks & Banks, 1995). Sackney (1988) reported several instruments that have been developed to measure school climate. The Pupil Control Ideology (PCI) measured the teacher-pupil relations while the Profile of Organizational Characteristics (POC), focused on leadership, communication, motivation, interaction-influence, decision making, goal setting, control and performance goals (Sackney, 1988). Halpin and Croft's (1963) research classified school climate into six types: open, autonomous, controlled, familiar, paternal, and closed. Some of these climates, such as open or familiar, are more conducive to the needs of diverse learners and teachers who implement culturally responsive pedagogy because they take into account the diversity and the background of their students. Students and teachers in such climates feel more comfortable and are given adequate support according to their needs.

Home and community support to teachers and students is another important aspect that influences the cultural beliefs of teachers. Teachers' understanding of diverse learners increases with increase in interaction of teachers with the students' home

culture. Only knowledge of diverse learners can help in constructing a foundation to use appropriate materials for instruction. In fact, Ladson-Billings (1990) found that parents of students of color expected good teachers to address a dual agenda which consists of helping their children achieve academic excellence and impart knowledge in such a way it does not alienate the children from their homes, community, or culture. In fact, they wanted their children to "hold their own in the classroom without forgetting their own in the community" (Ladson-Billings, 1990).

Epstein (1987) developed a model of “overlapping spheres of influence” which examines home-school relationships. The influence of families, schools, and communities is most effective when they have overlapping relationships (Epstein & Hollifield, 1996). In other words, when the school climate is completely alien to the home and community culture of the student, students of color are sometimes forced to lose their cultural identity and merge with the mainstream culture. This produces a mismatch of expectations between the family and the school. Epstein (1996) provided a model for different types and levels of parental involvement in their child’s education. They range from helping families establish home environment that supports their child’s role as a student, communication between home and school about their child’s progress and activities, parent volunteers in school activities, helping their child with home assignments and giving input with respect to the curriculum, playing an active role in the decisions of the schools, and collaborating and giving back to the community to strengthen all the spheres that influence the development and progress of children.

Factor III: Culturally Responsive Management

Culturally responsive management influences the cultural beliefs and attitudes of teachers towards diverse learners because of the difference between the learning styles of students from different backgrounds. For example, African American students use a lot of movements and energy in their classrooms which is often misdiagnosed as attention deficit, hyperactivity, or disruptive behavior (Gay, 2000). As Monroe (2005) noted,

> Empirical comparisons of cultural interaction styles indicate that teachers regularly interpret African American behaviors as inappropriate when the actions are not intended to be so. (p. 47)

In fact, Allen and Boykin (1992) named nine dimensions of Afro-cultural experience, which include spirituality, harmony, movement expressiveness, verve, communalism, expressive individualism, orality, and social time perspective. Research also shows that African Americans tend to have notably higher verve levels than their European American counterparts (Carter, Hawkins, & Natesan, forthcoming). However, this kind of movement-oriented, vervistic behavior is often dealt with using disciplinary actions by teachers. As a result, African Americans are often overrepresented in disciplinary settings across the nation (Gregory & Mosely, 2004).

The most common learning-centered traits possessed by African Amerrican students are *communalistic* learning and *vervistic* learning while the mainstream Eurocentric learning-centered traits are *individualistic* and *competitive* learning. Communalism represents the tendencies to prefer sharing ideas and materials along with helping others learn (Mbiti, 1970). It represents the tendencies to prefer sharing ideas

and materials along with helping others learn (Moemeka, 1998). Verve represents the tendency to remain energetic, intense, stimulating, and lively. Students with high levels of verve often multitask and shift focus among tasks rather than to focus on a single concern or a series of concerns in a rigid and traditional way (Boykin, 1983; Carter, Hawkins, & Natesan, In Press). Individualism represents focus towards working independently. Another aspect closely related to individualism is competition, which represents an orientation towards working competitively against others and the need to be the best at a given task. These are predominantly the characteristics exhibited by mainstream European American children (Spence, 1985).

In a study by Tyler, Boykin, and Walton (2006), teachers rated students' motivation and achievement to be higher if they displayed competitive and individualistic classroom behaviors, which are commonly exhibited by European American students, than if they displayed communal or vervistic behaviors which are commonly exhibited by African American students. Tyler, Boykin, and Walton (2006) noted that "few have examined directly teacher perceptions of culturally informed achievement behaviors of African American children" (p. 998). Their study found that the overall teachers' perceptions of student motivation and achievement were higher for mainstream cultural themes than for Afrocultural themes. Teachers also perceived a higher level of achievement for individualistic and competitive students than communalistic or vervistic students. Additionally, competitive and individualistic students were perceived as being more highly motivated than vervistic students.

As mentioned earlier, Sankofa, Hurley, Allen, and Boykin (2005) showed that children rated the high achievers differently, depending on the cultural orientation of the high achievers. They rated both the verve and the communalism high achievers higher than either the interpersonal competitive or individualism achievers. Therefore, Afro-cultural achievers received higher endorsement than the mainstream achievers. This further goes to prove the fact that the common perception of teachers that their students of color have disruptive classroom behavior is merely the lack of understanding of their learning styles. Teachers who can understand these differences are often able to convert and channel the energy of students of color to activities that are more productive. This, in turn, encourages the students to capitalize on what they bring to the classroom and become high achievers.

Factors V and VII: Cultural Sensitivity and Cultural Awareness

Cultural sensitivity and cultural awareness of teachers play an important role in their developing their attitudes about diverse learners. While cultural sensitivity addresses the knowledge of cultural similarities and differences without judging them (NMCHCCC, 1997), cultural awareness represents both cultural sensitivity and understanding (Adams, 1995). The two concepts are interrelated and can be considered the critical components of culturally responsive teaching. The primary aim of culturally-relevant teaching is to help students of color maintain their cultural identity and personality and achieve academic excellence. Ladson-Billings (1994) identified the attitudes of teachers towards students academically and culturally at-risk based on their tendency to seek excellence, improvement, or maintain status quo for their students and

their behavior to assume or shift the responsibility to others. Only teachers who strive for excellence and who share the responsibility of doing so with others can excel in culturally relevant teaching. She defined culturally-relevant teaching as the "antithesis of the assimilationist teaching" (p. 23).

Paley (1979) suggested teachers not to ignore color but instead to use it to acknowledge differences and build culturally relevant teaching upon it. Race and color have played such an extensive role in the history of the United States that ignoring the cultural differences would be the equivalent of dismissing the main features of the students' identities. However, equality must not be equated with sameness. Equality, instead addresses the different needs of different children by acknowledging their differences and dealing with them equitably. Ladson-Billings (1994) pointed out the eight main aspects of teachers who excel in culturally relevant teaching. They have: (1) high self-esteem and high regard for others, (2) see themselves as a part of the community, (3) believe all students can succeed, (4) help make connection between their community, national, and global identities, (5) see teaching as "digging knowledge out" of students, (6) demonstrate a connection with each of their students, (7) encourage a community of learners, and (8) encourage collaborative efforts. Culturally relevant teaching helps students connect their classroom experiences to their everyday lives.

Cultural awareness has been increasingly researched upon during the past few years. The Cultural Diversity Awareness Inventory (CDAI) developed by Henry (1986) has been modified and used by several researchers to measure the cultural awareness of teachers and the impact of multicultural education on their attitudes (e.g. Larke, 1990;

Milner, Flowers, Moore, Moore, & Flowers, 2003). Although courses about multicultural education improved the awareness of teachers in general, teachers still preferred to teach students whose cultural backgrounds were similar to that of the teachers. Most of these teachers perceived that the use of "non-standard English" in the classroom as inappropriate and the inclusion of parents of color in program planning as uncomfortable (Larke, 1990). However, a study by Swartz and Bakari (2003) reported that the tendency to be willing to work with students of color increased with increase in interactions between the teachers and students of color. Therefore, it can be concluded that more than teaching a multicultural course, hands-on interaction and experiences with students of color can help increase cultural awareness and sensitivity among teachers.

Factor VI: Curriculum and Instructional Strategies

The role of curriculum and instruction on student learning and pedagogical methods cannot be over emphasized. The learning styles, cognition, attitude, behavior, and personality of African Americans are different from that of the European Americans (Boykin, 1983; Irvine, 2003). However, the pedagogical practices, instruction, and materials in the American schools as of today are geared more towards the culture of European Americans, just as they have been in the past. Therefore, there is a gap between the popular pedagogical practices in schools and the learning styles of African American children and this is one of the main reasons for the academic achievement gap (Hale-Benson, 1986; Ramirez & Casteneda, 1974). Hilliard (1992) showed that African American students possess learning style characteristics that do not match the traditional

schools' analytical style of teaching. In fact, African American students find the school environment to be "unstimulating," "constraining," and "monotonous" (Boykin, 1983).

If the teachers are able to understand this and play to the strengths of their students, the achievement gap would narrow to a great extent. This shows that in order to be effective teachers, it is necessary for teachers to understand and accommodate the cultures, beliefs, and the backgrounds of their students. As Banks and Banks (1995, p. 157) noted, "Teachers who are skilled in equity pedagogy are able to use diversity to enrich instruction instead of fearing or ignoring it". There is no "one model fits all" approach when it comes to teaching a diverse class. Boykin and Toms (1985) found that African American students scored remarkably better on academic tasks when they were taught using vervistic methods.

According to several researchers, student academic outcomes are enhanced when aspects of their cultural background are present in classrooms (Boykin, 2001; Boykin, Albury, Tyler, Hurley, Bailey, & Miller, 2005). However, such themes usually are not prevalent in typical classroom settings especially in those attended by African American children from low income backgrounds. When elements of African American culture are incorporated into the curriculum and the instruction, African American children improve in performance, engagement, and motivation (Allen & Boykin, 1992; Bailey & Boykin, 2001; Boykin & Allen, 1988; Boykin, Allen, & Davis, 1997; Boykin & Cunningham, 2001; Dill & Boykin, 2000; Hurley, Allen, & Boykin, 2005). As Banks and Banks (1995, p.155) noted, "Equity pedagogy is most powerful when integrated with transformative curricula. "

The academic achievement of African American and Hispanic American students also increases when teachers use cooperative teaching strategies and activities (Aronson & Gonzalez, 1988). Cooperative learning has been shown to be an effective instructional technique by Cohen (1994) and Slavin (1983). However, Cohen and Roper (1972) also warned that, if used without an awareness of contextual issues such as differences among students, cooperative learning might reinforce stereotypes in the classroom. As Banks (1991b) stated, it is necessary to understand the differences between various groups in order to implement effective equity pedagogy. Only this knowledge can help in constructing a foundation to use appropriate materials for instruction (Ladson-Billings, 1990, 1994, 1995).

Banks (1991a) contended that the current curriculum in schools does not equip students to become reflective and critical citizens nor does it enable them to participate in the society to make it more equitable, democratic, and just. Curricular goals of racial and ethnic minorities have been labeled as "special interests" by mainstream scholars, which make people of dominant groups and people of color to think that mainstream curriculum is universal and caters to everyone. Thus, the mainstream curricula in schools and universities have led students to acquire beliefs that the values of dominant groups represent the values of the civic community. Such Eurocentric notions discourage and lower the sense of efficacy of students of color. Therefore, Banks (1991a) strongly recommended the inclusion and proper representation of

> events, concepts, and situations from the perspectives of the diverse cultural and racial groups within a society, including those that are politically and culturally dominant

> as well as those that are structurally excluded from full societal participation. (p. 127)

While transforming the curricula to include content about other ethnic groups, the dominant paradigms should not be used to select the content but rather content that portrays events and characters that added value to their communities must be included. Such transformative curricula help students to critically evaluate situations and become social critics who can make reflective decisions. In fact, he further identified the goal of a transformative curriculum as one that creates students who do reflective decision making and personal and civic action. Banks (1994b, p. 7) noted that, "the transformation approach brings content about the currently marginalized groups to the center of the curriculum."

Students exposed to multicultural materials enable them to interact more with students from other cultures and develop more positive racial attitudes (Slavin, 2001). As Sleeter (1995) noted, multicultural education is not simply integrating information about diverse cultural, ethnic, and racial groups into the mainstream curriculum. Multicultural education consists of five dimensions: content integration, the knowledge construction process, prejudice reduction, equity pedagogy, and empowering school culture and social structure (Banks, 1993a, 1994a, 1993b).

Factor VIII: Teacher Efficacy

According to Bandura (1995), "Perceived self-efficacy refers to beliefs in one's capabilities to organize and execute the courses of action required to manage prospective situations. Efficacy beliefs influence how people think, feel, motivate themselves, and act" (p. 2). There are four concepts that influence one's sense of efficacy, mastery

experiences, vicarious experiences, social persuasion, and physiological and emotional states. An example of mastery experience impacting self-efficacy is the encouragement and confidence one develops from successes. Vicarious experiences, such as seeing someone similar to one's self succeeding through perseverance, boosts the confidence that they possess the capabilities to achieve in similar arena. Social persuasion usually strengthens people's beliefs that they have what it takes to succeed. Such persuasive actions including talking to people increase their sense of efficacy. Physiological and emotional states such as moods, stress, tension, physical pain, and fatigue influence people's judgments of their own efficacy. Therefore, enhancing physical strain and reducing stress helps increase self efficacy. Efficacy plays a great role in self motivation. In other words, people's beliefs about what they can do influence what they will do.

Teachers with higher instructional efficacy provide successful experiences for their students. Classroom atmosphere, effective instruction, and the academic achievement of students depend upon the instructional and self-efficacy of teachers. Schools in which the staff collectively perceive themselves as ineffective in helping academically poor students improve their performance, create an atmosphere of low efficacy which also affects the teachers as a whole. Research has shown that in schools with staff that have higher self-efficacy, academically low-achieving students are motivated and are therefore able to achieve higher levels on standardized tests. Teachers with a high sense of instructional efficacy believe that with sufficient help from other resources such as parental intervention and appropriate instructional strategies, students with lower academic achievement can succeed. However, teachers with a low sense of

instructional efficacy believe that there is little that can be done about students with lower academic achievement (Bandura, 1997).

Teachers with high self-efficacy are more open towards learning new technological advances and pedagogical techniques. They plan their lesson plans with care, involve students in discussions, and are able to manage their classrooms with considerable ease (Saklofske, Michayluk, & Randhawa, 1988). Gibson and Dembo (1984) measured teachers' beliefs in their efficacy to motivate and educate academically low achieving students. Teachers with a high sense of efficacy devote more time towards academically-oriented activities, assist students with difficulty, and encourage their students. Therefore, these teachers create a sense of higher efficacy in their students as well. Teachers with a low sense of efficacy engaged in non-academically oriented activities and easily gave up on students with difficulty. The students of these teachers often have a lower sense of efficacy about themselves and their abilities.

Teacher efficacy is considered to be made of two dimensions, personal teaching efficacy and general teaching efficacy. Bandura (1997) defined personal teaching efficacy as the beliefs of a teacher about his/her ability to make a positive impact on the students. Teachers with high sense of personal teaching efficacy believe that all students are teachable. Teachers with a low sense of efficacy usually blame the students or their socioeconomic situations and factors beyond their control as the reason for student failure (Ashton & Webb, 1982; Pang & Sablan, 1998). A study by Gibson and Dembo (1984) showed that African American teachers are more likely believe that they can

bring about a positive change in the lives of students of color than their European American counterparts.

One of Ladson-Billings' (1994) suggestions for motivating change among teachers to make their teaching practices better place great emphasis on providing experiences and intervention to teachers that can help them understand the role of culture. A review of the existing literature shows that there is an immediate need to educate teachers to be culturally responsive and increase their awareness to help understand the needs of diverse learners. Such an understanding can help teachers be equitable, use pedagogical styles that would be more efficient in teaching students of color while maintaining their cultural identity, help narrow the academic achievement gap, and also teach the future of this country to be equitable, responsible, and righteous citizens.

Item Response Theory

Item Response Theory (IRT) or Latent Trait Measurement models were heralded as "one of the most important methodological advances in psychological measurement in the past half-century" (McKinley & Mills, 1989, p. 71). Their comparison to Classical Test Theory (CTT) is inevitable because CTT has been widely used. The most widespread argument about CTT is its inability to separate the test characteristics from the examinee characteristics (Henard, 2000). In fact, Hambleton and Swaminathan (1985) said that within CTT it is difficult to say whether an item is easy or difficult, because in CTT this depends on the abilities of the examinees. Conversely, it is also difficult to say if an examinee is

smart or not, because this depends on the difficulty level of the items used. Further, in CTT, item difficulties and person abilities are on different scales (Wright & Stone, 1979).

IRT overcomes these limitations and helps the researcher build items free from examinee and test item biases (Henard, 2000; Wright & Stone, 1979). IRT transforms the item difficulties and person abilities into estimates on a single scale which is theoretically both "person-free" and "item-free" (Cantrell, 1999). There are several models in IRT based on the number of parameters used. IRT has two main postulates, mainly the latent traits and the Item Characteristic Curve (ICC) (Cantrell, 1999). Latent traits (or simply traits or abilities) measure the performance of an examinee on a test item. An ICC is a frequency polygon or ogive representing the relationship between the item performance and the examinee's set of traits that determine examinee performance (Cantrell, 1999; Hambleton & Swaminathan, 1985).

There are potentially as many as three parameters that determine the likelihood of an item being answered correctly. They are (a) the item discrimination parameter, "a", (b) the item difficulty or the location parameter, "b", and (c) the guessing parameter, "c" (Lord & Novick, 1968). The item discrimination parameter, as the name indicates, represents how well the item can distinguish persons with different trait levels. For example, if the item discrimination parameter is high, this means that for a small change in ability (or attitude level), the probability of endorsing the item is high. This parameter is represented by the slope of the ICC.

The item location parameter directly represents the endorsement level of the item. The position of the ICC represents the location parameter and the Y intercept of the curve

represents the guessing parameter, which corresponds to the probability of endorsement of a person with least or no ability (or attitude level). These three parameters combined with the person's ability (or attitude level) give rise to the ICC.

There are special cases of the IRT model where one or more of the parameters are dropped. A two-parameter model consists of only the item discrimination and item difficulty parameters, omitting the guessing parameter. In other words, a two-parameter model is the special case of the three-parameter model with the guessing parameter always set to zero. Similarly, in a one-parameter model both guessing and item discrimination parameters are not included and hence not modeled (Hambleton & Swamination, 1985). In other words, a one-parameter model is a special case of the three-parameter model where the guessing parameter is zero and all the slopes or discrimination parameters are equal and hence only the influence of item difficulty is considered.

The three-parameter logistic (3-PL) model, for dichotomous items, defines the probability of a positive response to an item i ($x_i = 1$) as

$$T\left(x_i = 1|\theta\right) = c_i + (1 - c_i)\frac{1}{1 + \exp\left[-a_i\left(\theta - b_i\right)\right]}, \tag{48}$$

where a_i is the item discrimination (or slope) parameter, b_i is the item location parameter, and c_i is the guessing parameter (Birnbaum, 1968).

In a two-parameter logistic (2-PL) model, for dichotomous items, the guessing parameter is set to zero. Therefore, the probability of a positive response to an item i ($x_i = 1$) as

$$P\left(x_i = 1|\theta\right) = \frac{1}{1 + \exp\left[-a_i\left(\theta - b_i\right)\right]}, \tag{49}$$

Similarly, in a 1-PL model, the discrimination parameter, *a*, is set to 1. Two-parameter MLIRT models were used in this study instead of 1-PL MLIRT models to demonstrate the advantages of including the discrimination parameter in the model as opposed to simply using the threshold parameters alone. Similarly the 2-PL MLIRT model was preferred over the 3-PL MLIRT model because when measuring the attitudes of people, the guessing parameter (the third parameter in the 3-PL model) seems redundant because the respondents are usually aware of what their attitudes are about a certain construct.

Multilevel IRT Models

The most recent development in the field of IRT is the concept of Multilevel IRT models (MLIRT). When the effects of multilevel covariates on a latent trait need to be estimated, IRT and Multilevel Modeling can be combined. This amalgamation of the two models allows us investigate and analyze the covariates that affect the person abilities instead of simply estimating the latent traits (Maier, 2001). This merger also paves way to modeling the abilities over time when repeated observations are made, or across various raters, or simply for people belonging to a certain group versus another.

Advantages of MLIRT

Why would a researcher prefer MLIRT to other IRT estimation techniques, given that the estimates yielded by MLIRT are comparable to the other IRT estimation techniques for both dichotomous (Kamata, 1998) and polytomous items (Williams, 2003)? In many IRT techniques, the item and person parameters are estimated simultaneously. This gives rise to the "Neyman-Scott problem", which is the inconsistency in the estimates of item and person parameters when they are estimated

simultaneously (Neyman & Scott, 1948). MLIRT allows the item parameters to be treated as fixed and person abilities as random parameters, thereby avoiding the Neyman-Scott problem (Kamata, 2001).

Multilevel formulation of IRT facilitates the modeling of multiple-group IRT models (Bock & Zimowski, 1997) and thereby its special cases, such as group-level IRT model (Mislevy, 1983; Mislevy & Bock, 1989), item parameter drift model (Bock, Muraki, & Pfeiffenberger, 1988), and the duplex design model (Bock & Mislevy, 1989). Therefore, the effects of the variables such as person or group-characteristics can be evaluated. In the two-level analysis, when person characteristics such as gender, are taken into account, the effect of those characteristics can be estimated. Similarly, the three-level analysis, when group membership and the hierarchical structure of the data are taken into account, estimates the effects of group-level and person-level abilities, the interaction effects of person characteristics and group membership, and the estimate of person-level effects across groups (Kamata, 2001; Williams, 2003). This provides additional information about the parameter estimates at each level of the model, thereby avoiding the need to perform separate analyses (Adams, Wilson, & Wu, 1997; Kamata, 1998).

An often used polytomous IRT model is Samejima's (1969, 1997) graded response model (GRM), a generalization of the 2PL model that permits estimation of multiple b_{ij} parameters per item (j from 1 to n-1) associated with n response categories. The formula for a GRM trace line is:

$$P(x_i = j|\theta) = \frac{1}{1+\exp[-a_i(\theta - b_{ij})]} - \frac{1}{1+\exp[-a_i(\theta - b_{ij+1})]} \quad (50)$$

which states that the probability of responding in category *j* is the difference between a 2PL trace line for the probability of responding in category *j* or higher and a 2PL trace line for the probability of responding in category *j*+1 or higher. Samejima's model was used for this study.

Instrument

The teachers' perceptions of Cultural Awareness and Beliefs Inventory (CABI) survey was developed by Webb-Johnson and Carter (2005) and administered to teachers in from various schools in an urban area located in the Southeastern part of the United States in 2006 (Appendix C). It consists of 45 items measuring the cultural perceptions of teachers on various scales such as school climate, home and community support, teacher efficacy, curriculum and instructional strategies, belief system, and cultural awareness. The demographics and background information include the school for which they work, gender, ethnicity, teaching experience, grade level, certification, and the type of degree obtained by the teachers. This instrument was administered in the form of a paper and pencil survey where the respondents had to mark their answers on a Scantron sheet. The advantage of using this method was the ease of encoding the data especially because of the large sample size. However, a significant amount of teachers (about 300) did not mark their answers properly on the sheets (by either not using a 2H pencil or not shading the circles properly) and therefore some valuable data was lost.

Methodology

Prior Analyses

Validity and reliability were already established for scores on this instrument as a part of another study (Walter-Roberts, Natesan, & Carter, manuscript under preparation). Factor analysis was performed and the factors were identified. The internal consistency measured by Cronbach's alpha was found to be 79.9% with no noteworthy change in reliability if any item were to be deleted. However, when factor analysis was conducted to establish convergent and divergent validity, 10 items were deleted from the instrument because they did not have a factor pattern/structure coefficient of more than 0.4 on any factor (Thompson, 2004). These included questions 16, 18, 24, 29, 33, 36, 43, 44, 45, and 54. Eight factors were obtained and these eight factors explained about 44% of the variance. Based on the analysis, the factors were named as (a) Teacher Beliefs (b) School Climate, (c) Culturally Responsive Management, (d) Home and Community Support, (e) Cultural Sensitivity, (f) Curriculum and Instructional Strategies, (g) Cultural Awareness, and (h) Teacher Efficacy. The factor coefficients of the items on the factors and the reliability of each individual scale can be seen in Tables C1 and C2 respectively in Appendix C. The same dataset from the factor analysis study was used in the present study.

MLIRT Analyses

As mentioned in the literature review of study-1, IRT models can be viewed as multilevel models by considering the item difficulty or location as the first level variable and the latent trait as the second level variable (Kamata, 1998). The two main

assumptions of several IRT models are *unidimensionality* and *local independence* of items. Unidimensionality refers to measurement of a single latent trait. Although multidimensional models have been developed (Reckase, 1997; Reckase & McKinley, 1991), the current study will focus on unidimensional models only. The assumption of local independence states that the responses to two different items are independent of each other (Hambleton & Swaminathan, 1985). Items belonging to a single factor were considered unidimensional in the present study. It was reasonable to assume this because both convergent and divergent validities were established for the instrument which indicates that the items that belonged to a particular factor had high correlations among themselves and did not have a high correlation with any other factor.

Estimation of parameters for each factor that impacts the cultural beliefs of teachers were performed using GLLAMM (Generalized Latent Linear and Mixed Models), a user-written program in Stata (Rabe-Hesketh, Skrondal, & Pickles, 2004b). This indicated the endorsement rate and response consistency of the items by the teachers for each factor. GLLAMM was conducted for the initial analysis without any predictor variable (without the *geqs* option) to simply identify the parameters in general.

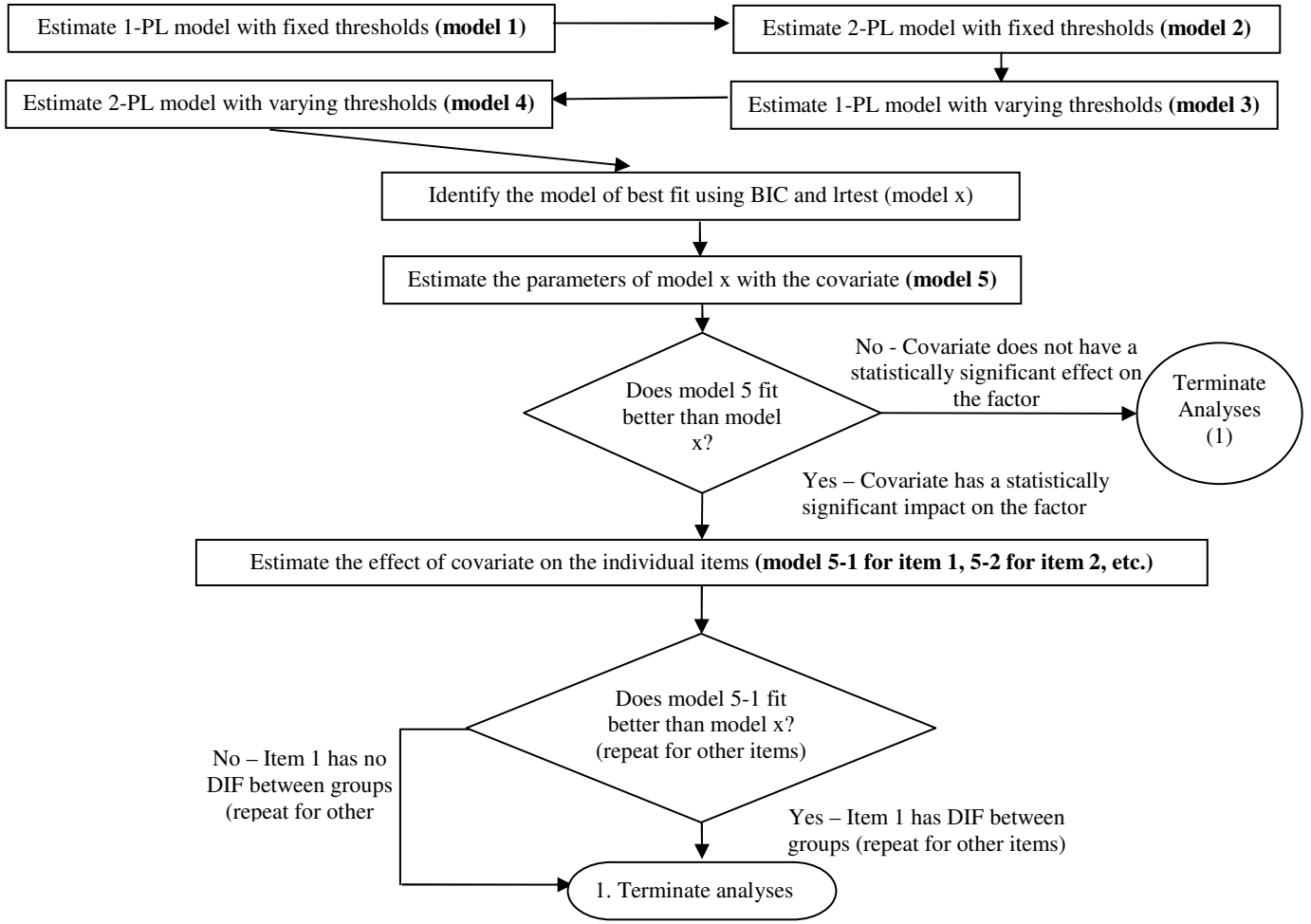


***Figure* 3. Decision Tree for MLIRT and DIF analyses**

Figure 3 shows the decision tree that can be used to determine the model of best fit for the MLIRT analyses and the DIF analyses. In order to find the overall and differential effects of the covariates on the factor and the items, GLLAMM was conducted separately with ethnicity, gender, level taught, and teaching experience as predictor variables. The purpose of this analysis was two-fold. The first was to investigate if there are any differences between or among teachers belonging to different ethnic groups, gender, level taught or teaching experience with respect to their cultural beliefs and awareness. The second was to identify if there is any bias in the items with respect to measuring the cultural attitudes between or among teachers belonging to different groups demographically. This was done by evaluating two models, one that measures the effect of the predictor variable (in this case, group membership) on the factor, and the other that measures the direct effect of the predictor variable on the individual items, in addition to the effect on the entire factor, and then finding which model the data fits better. The graphical representation of the 2-PL MLIRT model for the teacher beliefs factor with a covariate is presented in Figure 4.

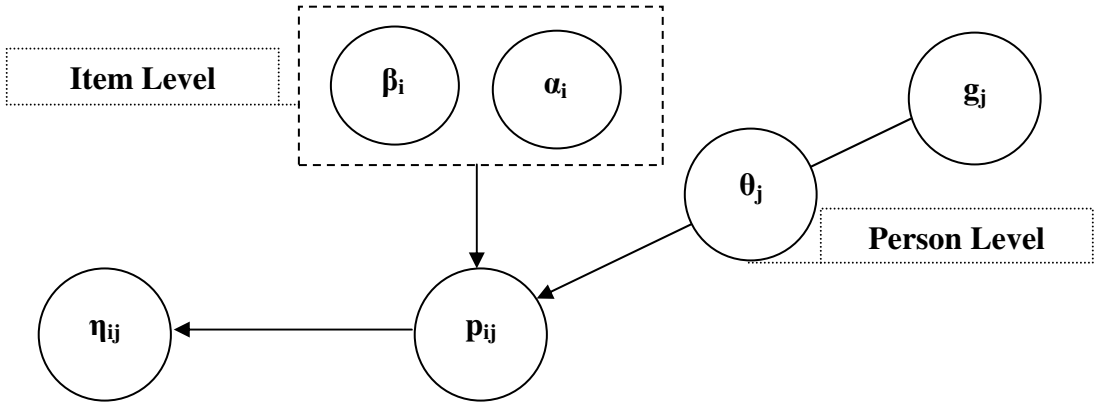


Figure 4. **Graphical Representation of the 2-PL MLIRT Model for Factor I-Teacher Beliefs with a Covariate**
θ_j is the latent trait (Teacher Beliefs about African American children) of person j; λ_j is the gender (covariate) of teacher beliefs of person j; β_i is the slope (discrimination parameter) of item i; α_i is the threshold parameter of item i; p_{ij} is the probability of person j choosing the category η_{ij} for item i

Bayesian Information Criterion (BIC)

Although Rabe-Hesketh, Skrondal, and Pickles (2004b) suggested investigating the statistical significance using log likelihood ratio test only, it is highly necessary to include the *Bayesian Information Criterion (BIC)* estimates as well in order to make decisions about model fit. This is because the likelihood ratio tests are known to be a mixture of chi-squares and not a single chi-square distribution in several circumstances (Weiss, 2005) in addition to not being robust to large sample sizes (Powers & Xie, 2000; Raftery, 1986, 1995). Therefore, the BIC statistic was calculated using the G-square value from the log likelihood test (G^2=*-2 x* ($L_1^2 - L_2^2$); where L_1^2 and L_2^2 are the log likelihoods of the previous and the new models) using the following formula:

$$BIC = G^2 - (DF \times logn)$$

where *DF* is the degrees of freedom (number of items-1) and n is the sample size (Raftery, 1986, 1995). BIC was chosen over *AIC (Akaike Information Criterion)* for the present study because "BIC more than AIC tends to pick the null model when the null model is in fact correct" for large samples (Weiss, 2004). The advantage of using BIC over log likelihood can be seen when interpreting the results of the analyses in the present study. A negative BIC between two models indicates that the previous model has better fit than the current model. The use of BIC can be considered analogous to reporting effect sizes in classical test theory. Therefore, all the tables that report the chi-square values and the p-values for the chi-square values also report the BIC values between two models and BIC was chosen over p-values to identify the model of best fit.

Results

Population

The CABI study was conducted in an urban school district located in an urban school district in the southeastern part of the United States. This urban school district is situated in the third most populous county in the state and has been rated as the second fastest growing among the ten most populous counties with over 3 million residents in the United States (Fast Facts, 2006). The archival data for the present study was collected in an urban school district located in 111 square miles in southeastern Texas. Sixty-six campuses in this urban school district consist of 4,537 teachers that serve 56,255 students.

The district consisted of 42% European American teachers (n=1885), 28% African American teachers (n=1214), 12% Hispanic American teachers (n=563), and 2% teachers belonging to other ethnicities (n=69) in the year 2004 (Texas Education Agency [TEA], 2004). The student body consists of 60% Hispanic Americans (n=33,918), 32% African Americans (n=17,836), 6% European Americans (n=3,215), 2% Asian/Pacific Islander American (n=1,238), and 0.08% Native American (n=48) as shown in Table 4 (TEA, 2004).

Table 4
Ethnic Composition of the Urban School District's Student and Teacher Populations

	Students		Teachers	
Ethnicity	**N**	**Percentage**	**N**	**Percentage**
Hispanic American	33,918	61%	563	12%
African American	17,836	32%	1,214	28%
European American	3,215	5%	1,885	42%
Asian/Pacific Islander	1,238	2%	(Other) 69	2%
Native American	48	.08%		
Total	***56,255***	***100%***	***3,731***	***100%***

The target population for the present study was in-service teachers instructing Pre-Kindergarten through grade 12 students in an urban public school district. Fifty-four individual campuses with approximately 4000 elementary and secondary classroom teachers volunteered to participate in the study. As mentioned in chapter III, instrument

scores were validated as part of a previous study (Walter-Roberts, Natesan, Carter, & Webb-Johnson, Under Preparation).

Sample

Out of the 3,731 teachers, 1,046 answered the survey giving a response rate of about 28%. Although the return rate was low, the sample size is quite high (1046) and adequate to perform some advanced statistical analyses such as MLIRT analysis. Of the 1046, 79.98% (n=837) were females and 20.08% (n=209) were males. Among the respondents, 39.75% (n=584) were European Americans, 27.71% (n=407) African Americans, 19.33% (n=284) Hispanic Americans, 2.25% (n=33) Asian/Pacific Islanders, 2.25% (n=33) Native Americans, and 8.71% (n=128) belonged to other ethnic groups. However, for MLIRT analysis, missing data has to be deleted from the database. Therefore, after the missing data was deleted and groups were combined, there were 33.62% (n=351) European Americans, 26.05% (n=272) African Americans, 23.95% (n=250) Hispanic Americans, and 16.38% (n=171) teachers from other ethnic groups. An interesting observation to be noted here is the overrepresentation of Hispanic American and under representation of European American teachers in the sample after the missing data was deleted. This shows that European American teachers refrained from answering some questions on their cultural beliefs and awareness about teaching African American students compared to other teachers and this has affected the overrepresentation of Hispanic American teachers. The statistical effect of this on the estimates of the item parameters is minimal because the proportion of group sizes in ethnicity does not affect the estimates as shown in the simulation study. Additionally, the

correlation between the predictor (ethnicity) and the attitude levels of teachers does not affect the estimates of item parameters as well.

The database consisted of teachers with teaching experience ranging from 1 month to more than 10 years. About 15.95% (n=208) had a teaching experience ranging from 1 to 11 months, 21.63% (n=282) had 1-3 years, 25.15% (n=328) had 4-6 years, 17.56% (n=229) had 7-9 years, and 19.71% (n=257) had 10 or more years of teaching experience. Teachers with a teaching experience of 3 years or less were considered as novices and 4 years or more as veterans. Therefore, after missing data was deleted and groups combined, there were 38.23% (n=398) novice teachers and 61.77% (n=643) veteran teachers. Although the group sizes for gender and teaching experience were not comparable, the effect of these groups on the different latent traits representing the cultural awareness and beliefs of teachers were estimated. This gives a general idea about the impact of these groups on the latent traits of interest. The demographics of the sample before and after combining the groups are shown in Figures 5 to 7. The results of the study are discussed below. Instead of discussing the results in the chronological order of the research questions, this section focuses on the answering the research questions by factor. Therefore, each section deals with the parameters (question 3) and also the direct and the differential effect of the covariates (gender, ethnicity, and teaching experience) on the items in each factor.

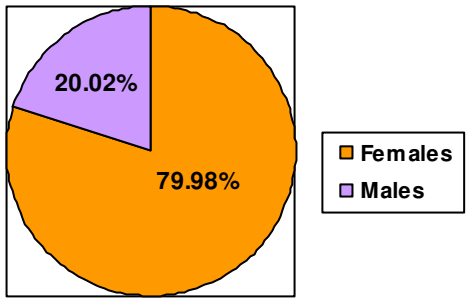


Figure 5. Gender of the Respondents

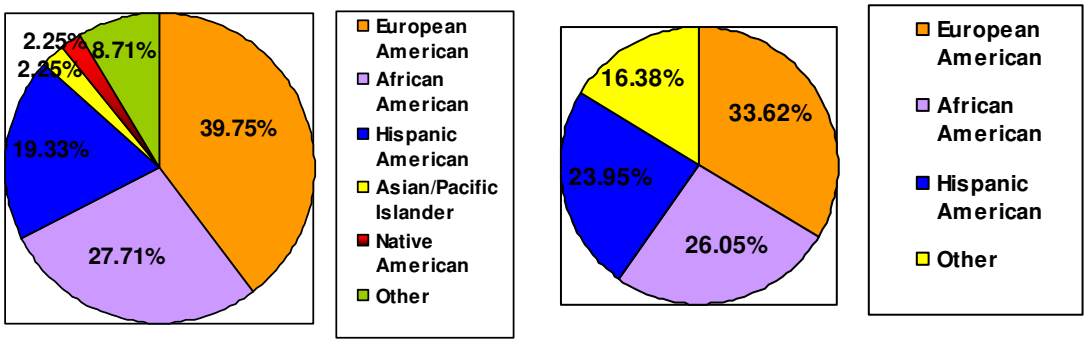


Figure 6. Ethnicity of the Respondents Before and After Combining Groups

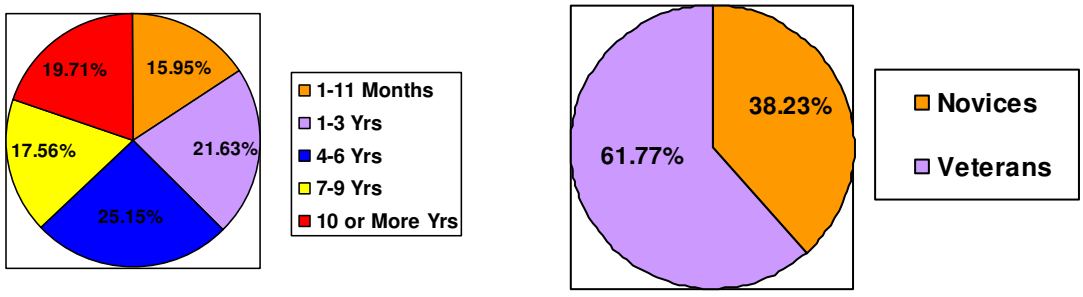


Figure 7. Teaching Experience of the Respondents Before and After Combining Groups

Factor I: Teacher Beliefs about Teaching African American Students

The teacher beliefs factor consisted of 8 items as shown in Table 5. Multilevel Item Response theory was applied to this factor, with the latent trait of interest (dependent variable) being beliefs of teachers in teaching African American students. The analysis was first conducted to investigate the qualities of the items and how much they contribute to teacher beliefs about African American students. Then the effects of gender, ethnicity, and teaching experience of the teachers on their beliefs were estimated both for differential and the overall effect of these covariates on teacher beliefs.

The item location (threshold) and slope (discrimination) parameters for the items measuring teacher beliefs for various models are shown in Table 6. After considering the chi-square and the BIC values between the models, the 2-PL IRT model with varying thresholds was found to be the model of best fit for factor I. The difference between the BICs of the 2-PL IRT model with varying thresholds and the 1-PL IRT model with varying thresholds was found to be 129.41 which shows that the former model fits much better than the latter. The BIC is a more appropriate statistic than the chi-square statistic because it is more robust given sample sizes (Powers & Xie, 2000; Raftery, 1986, 1995).

Table 5
Items in Teacher Beliefs Factor

Question No.	Item
30	I believe African American students consider performing well in school as “acting White”.
31	I believe African American students have more behavior problems than other students.
32	I believe African American students are not as eager to excel in school as White students.
34	I believe students who live in poverty are more difficult to teach
35	I believe African American students do not bring as many strengths to the classroom as their White peers.
38	I believe I would prefer to work with students and parents whose cultures are similar to mine.
42	I believe I have experienced difficulty in getting families from African American communities involved in the education of their students.
52	I believe students from certain ethnic groups appear lazy when it comes to academic engagement.

Table 6

Item Thresholds and Discrimination Parameters for Models 1 through 5-5 (Factor 1)

Items	Parameters	1-PL (model 1)	2-PL (model 2)	1-PL (model 3)	2-PL (model 4)	gender cov. (model 5)	Ethnicity cov. (model 6)	Exp. cov. (model 7)
Item 30	Threshold 1-2	-2.300	-2.313	-2.062	-1.933	-2.120	-2.270	-2.050
	Threshold 2-3	-1.119	-1.115	-1.155	-1.073	-1.260	-1.400	-1.190
	Threshold 3-4	0.490	0.519	0.468	0.446	0.260	0.109	0.328
	Discrimination	0.854	0.748	0.854	0.658	0.657	0.637	0.656
Item 31	Threshold 1-2	-2.740	-2.682	-1.707	-2.050	-2.427	-2.789	-2.290
	Threshold 2-3	-1.559	-1.484	-0.620	-0.779	-1.156	-1.492	-1.019
	Threshold 3-4	0.049	0.150	0.801	0.951	0.573	0.259	0.710
	Discrimination	0.854	1.271	0.854	1.339	1.331	1.331	1.336
Item 32	Threshold 1-2	-2.325	-2.239	-2.281	-3.343	-3.850	-4.267	-3.670
	Threshold 2-3	-1.144	-1.041	-1.157	-1.721	-2.232	-2.637	-2.049
	Threshold 3-4	0.464	0.592	0.561	0.763	0.251	-0.174	0.436
	Discrimination	0.854	1.229	0.854	1.822	1.807	1.758	1.821
Item 34	Threshold 1-2	-2.828	-2.836	-1.788	-1.668	-1.856	-2.013	-1.786
	Threshold 2-3	-1.647	-1.638	-0.501	-0.463	-0.650	-0.796	-0.581
	Threshold 3-4	-0.038	-0.004	-0.921	0.863	0.675	0.523	0.743
	Discrimination	0.854	0.776	0.854	0.658	0.657	0.643	0.657
Item 35	Threshold 1-2	-2.017	-1.940	-2.589	-2.965	-3.300	-3.566	-3.190
	Threshold 2-3	-0.836	-0.742	-1.388	-1.595	-1.933	-2.193	-1.817
	Threshold 3-4	0.773	0.892	0.201	0.195	-0.145	-0.423	-0.024
	Discrimination	0.854	1.116	0.854	1.209	1.199	1.159	1.212
Item 38	Threshold 1-2	-2.431	-2.449	-2.249	-2.110	-2.301	-2.454	-2.229
	Threshold 2-3	-1.250	-1.252	-1.143	-1.065	-1.256	-1.398	-1.185
	Threshold 3-4	-0.359	0.382	0.754	0.706	0.516	0.361	0.585
	Discrimination	0.854	0.664	0.854	0.675	0.673	0.649	0.669
Item 42	Threshold 1-2	-3.179	-3.191	-1.500	-1.376	-1.539	-1.661	-1.482
	Threshold 2-3	-1.998	-1.993	-0.161	-0.137	-0.301	-0.418	-0.242
	Threshold 3-4	-0.389	-0.359	1.306	1.186	1.020	0.887	1.082
	Discrimination	0.854	0.672	0.854	0.585	0.580	0.552	0.584

Table 6 (continued)

Items	Parameters	1-PL (model 1)	2-PL (model 2)	1-PL (model 3)	2-PL (model 4)	gender cov. (model 5)	Ethnicity cov. (model 6)	Exp. cov. (model 7)
Item 52	Threshold 1-2	-2.376	-2.392	-2.459	-2.342	-2.547	-2.697	-2.474
	Threshold 2-3	-1.195	-1.194	-1.109	-1.054	-1.260	-1.408	-1.186
	Threshold 3-4	0.414	0.440	0.653	0.624	0.417	0.250	0.492
	Discrimination	0.854	0.718	0.854	0.736	0.731	0.698	0.734
Model	Log Lkhd*	-8687.9	-8594.3	-8644.4	-8515.1	-8511.2	-8480.5	-8513.88
	BIC¥		166.05	79.06	129.41	-13.38(***N***)	48.07(***Y***)	-18.59(***N***)
	Random cov. Effect	-	--	--	--	-0.155	-0.076	-0.073
	P**	--	*0.00*	*0.00*	*0.00*	*0.00*(***Y***)	*0.00*(***Y***)	*0.11*(***N***)

Note: * Log Lkhd represents the Log Likelihood of the model; BIC¥ represents the BIC of current model over previous model; ** P represents the probability that the previous model fits better than the current model

For simplistic representation, the threshold parameter between strongly disagree and disagree will be represented as 1-2 threshold, between disagree and agree will be represented as 2-3 threshold, and between agree and strongly agree as 3-4 threshold for the rest of this study. Items 32, 35, and 52 had the lowest 1-2 threshold parameters, which show that teachers are less likely to strongly disagree with these items than the rest of the items in this factor. Items 31, 32, and 34 also had the highest 3-4 threshold parameters showing that teachers are less likely to strongly agree with these items compared to the other items. Therefore, one would expect the respondents to be more likely to choose both the extreme and the middle categories. However, as we can see from the item characteristic curves in Figure 8, that is not the case for some of the items. This can be attributed to the fact that this is a 2-PL model and therefore, only when the threshold parameters along with the discrimination parameters are considered in the interpretation of data, will there be a clearer understanding about the items.

The discrimination parameters for the 8 items ranged from (1.822 to 0.585). Item 32 (“I believe African American students are not as eager to excel in school as White students”) had the highest discrimination parameter, which means that this item contributed the most towards the latent trait, teacher beliefs. Items 31 and 35 had high discrimination parameters as well. Items 42, 30, 38, and 52 had low discrimination parameters. When the items have low discrimination parameters, these items are almost equivalent to being dichotomous, as can be seen from the item characteristic curves in Figure 8.

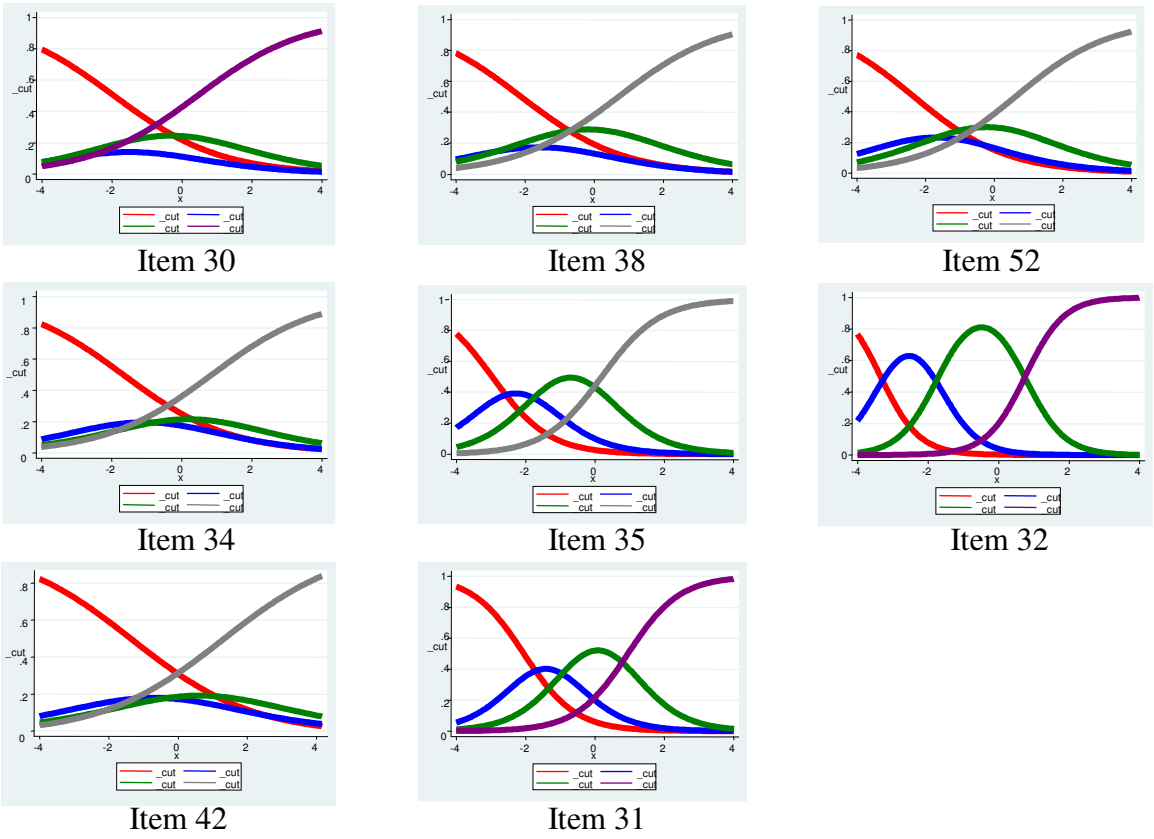


***Figure* 8. Item Characteristic Curves for Factor I – Teacher Beliefs About Teaching African American Students**

Therefore, our expectation about the middle categories for item 52 cannot just be based on the threshold values because it can be seen that having a low discrimination parameter suppresses the model categories. This also shows the importance and advantages of the 2-PL model over the 1-PL model. The slope parameters in IRT can be considered as the counterpart of beta-weights or factor pattern coefficients. Therefore, a comparison of the discrimination parameters of the items with their respective factor

pattern coefficients on the factor (from Table C1) can be conducted to better understand the similarities and differences between the two statistics. Some of these values are also consistent with the factor coefficients of the items on the factor in the previous study. Items 31, 32, and 35 had the highest factor pattern coefficients (0.785, 0.81, and 0.745 respectively) as well as high discrimination parameters. However, item 38 which had the lowest factor pattern coefficient (0.444) did not have the lowest discrimination parameter, indicating that the results shown by factor analysis are not identical to the results shown by item response analysis.

Gender, ethnicity, and teaching experience were included in the model to find the effect of these covariates on the items (both direct and random). The p-value and the BIC values between the 2-PL model with varying thresholds and the model with the covariate of interest were both considered to make a decision about the effect of the covariates on the factor. Gender and teaching experience did not have a statistically significant impact on teacher beliefs. However, ethnicity had a significant influence on teacher beliefs as shown in Table 6 (random effect/beta co-efficient = -0.076; BIC = 48.07). In order to find the differential effect of ethnicity on the individual items, post hoc analyses were conducted. The results of the post hoc test are shown in Table 7.

Through the post hoc analyses, the random covariate effect on the items and the direct effect of the covariate on each individual item were identified. Again, the BIC and the p-values provided the rule of thumb to decide the items that had differential item functioning according to ethnicity. As shown in Table 7, ethnicity had a statistically significant overall impact on teacher beliefs, but there was no differential effect on 7 of the 8 items. Only item 42 ("I believe I have experienced difficulty in getting families from African American communities involved in the education of their students") had differential item functioning across ethnic groups.

The trace lines (item characteristic curves) for the 4 different ethnicities taken two at a time (European American, African American, Hispanic American, and Other) for the 8 items are shown in Figures 9 – 16. However, decision about which ethnic groups were similar to/different from other ethnic groups cannot be made from these trace lines because they are not controlled for random covariate effect on the items. Instead, they give a general idea of the performance of various groups on these 8 items.

Table 7
Item Thresholds and Discrimination Parameters for Models with Statistically Significant Covariates (Factor I)

Covariate		**Item 30**	**Item 31**	**Item 32**	**Item 34**	**Item 35**	**Item 38**	**Item 42**	**Item 52**
Ethn.	Threshold 1-2	-2.309	-3.055	-4.260	-2.106	-3.511	-2.459	-1.392	-2.513
	Threshold 2-3	-1.449	-1.772	-2.640	-0.901	-2.140	-1.414	-0.123	-1.213
	Threshold 3-4	0.069	-0.002	-0.169	0.424	-0.351	0.355	1.223	0.484
	Discrimination	0.625	1.282	1.750	0.621	1.174	0.644	0.628	0.752
Model	Log Lkhd*	-8480.336	-8470.297	-8480.540	-8479.120	-8479.880	-8480.540	-8466.500	-8473.560
	BIC¥	-20.719	-0.639	-21.129	-18.289	-19.809	-21.129	6.950	-7.159
Random Cov. Effect		-0.074	-0.067	-0.076	-0.074	-0.078	-0.076	-0.080	-0.079
Cov. Effect on Item		-0.010	-0.083	0.002	-0.026	0.022	-0.001	0.084	0.060
	P**	0.515	0.000	0.919	0.091	0.249	0.921	0.000	0.000
	Presence of **DIF**	**NO**	**NO**	**NO**	**NO**	**NO**	**NO**	**YES**	**NO**

Note: * Log Lkhd represents the Log Likelihood of the model; BIC¥ represents the BIC of current model over previous model; ** P represents the probability that the previous model fits better than the current model

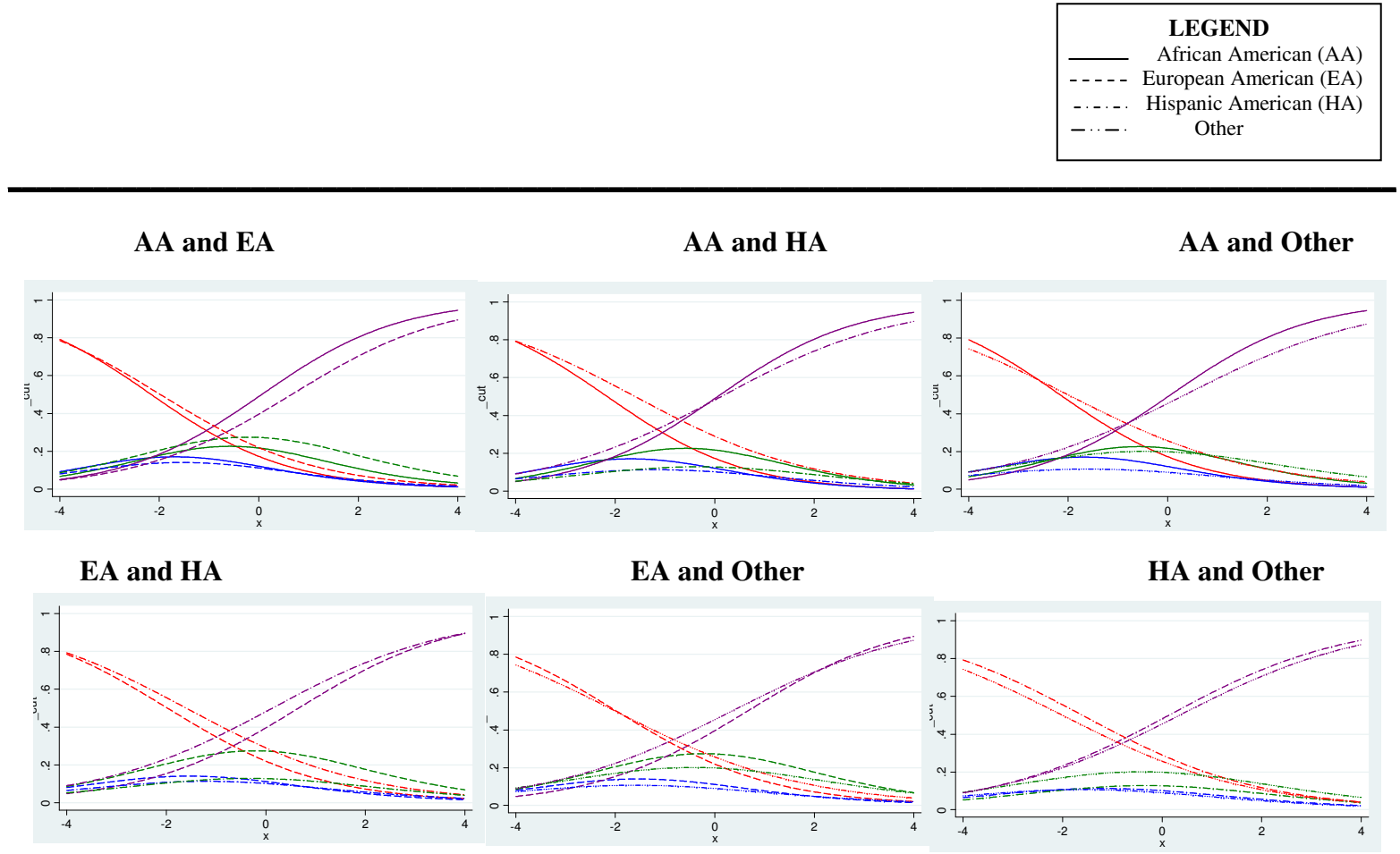


***Figure* 9. Item Characteristic Curves for Item 30 by Ethnicity**

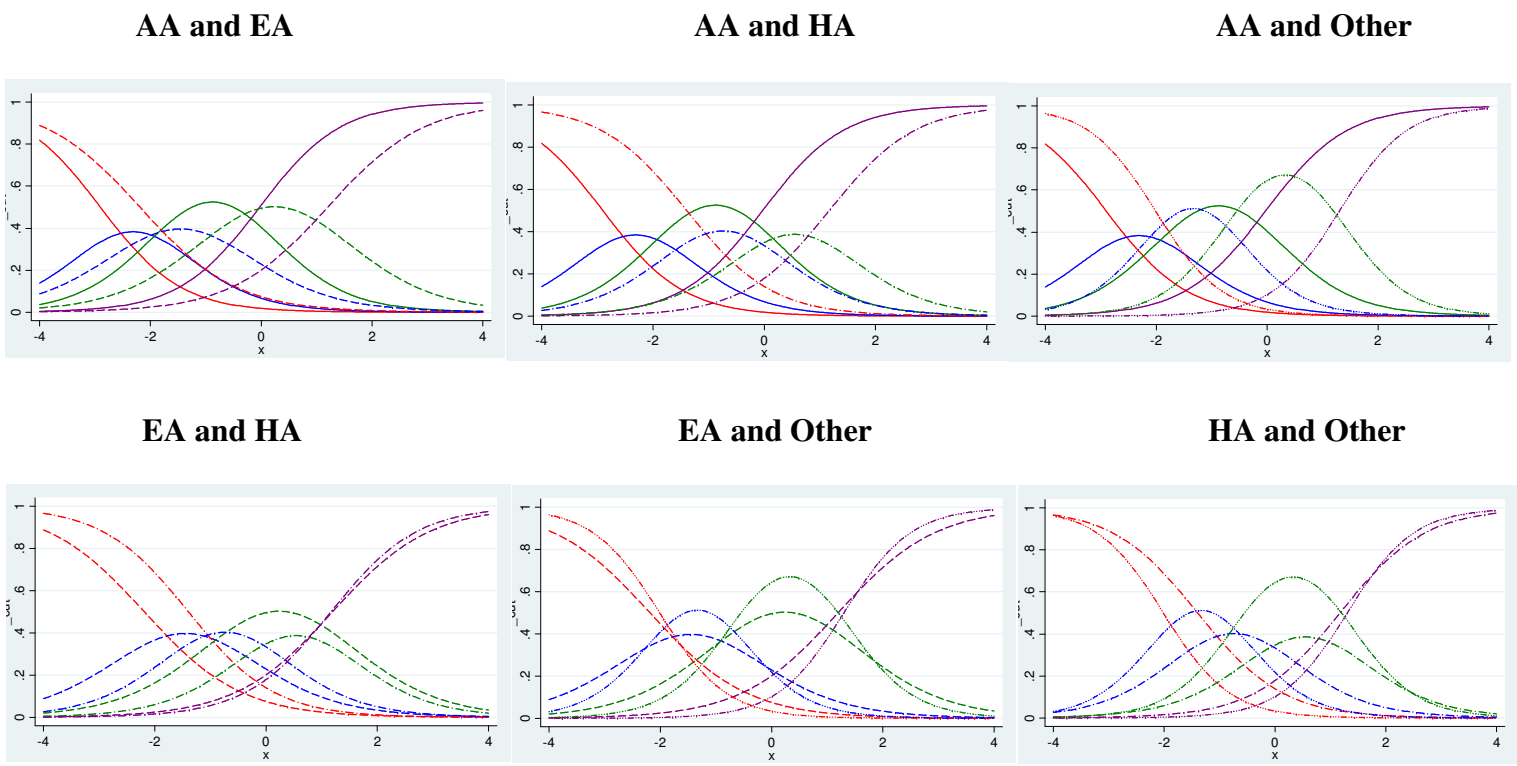


Figure **10. Item Characteristic Curves for Item 31 by Ethnicity**

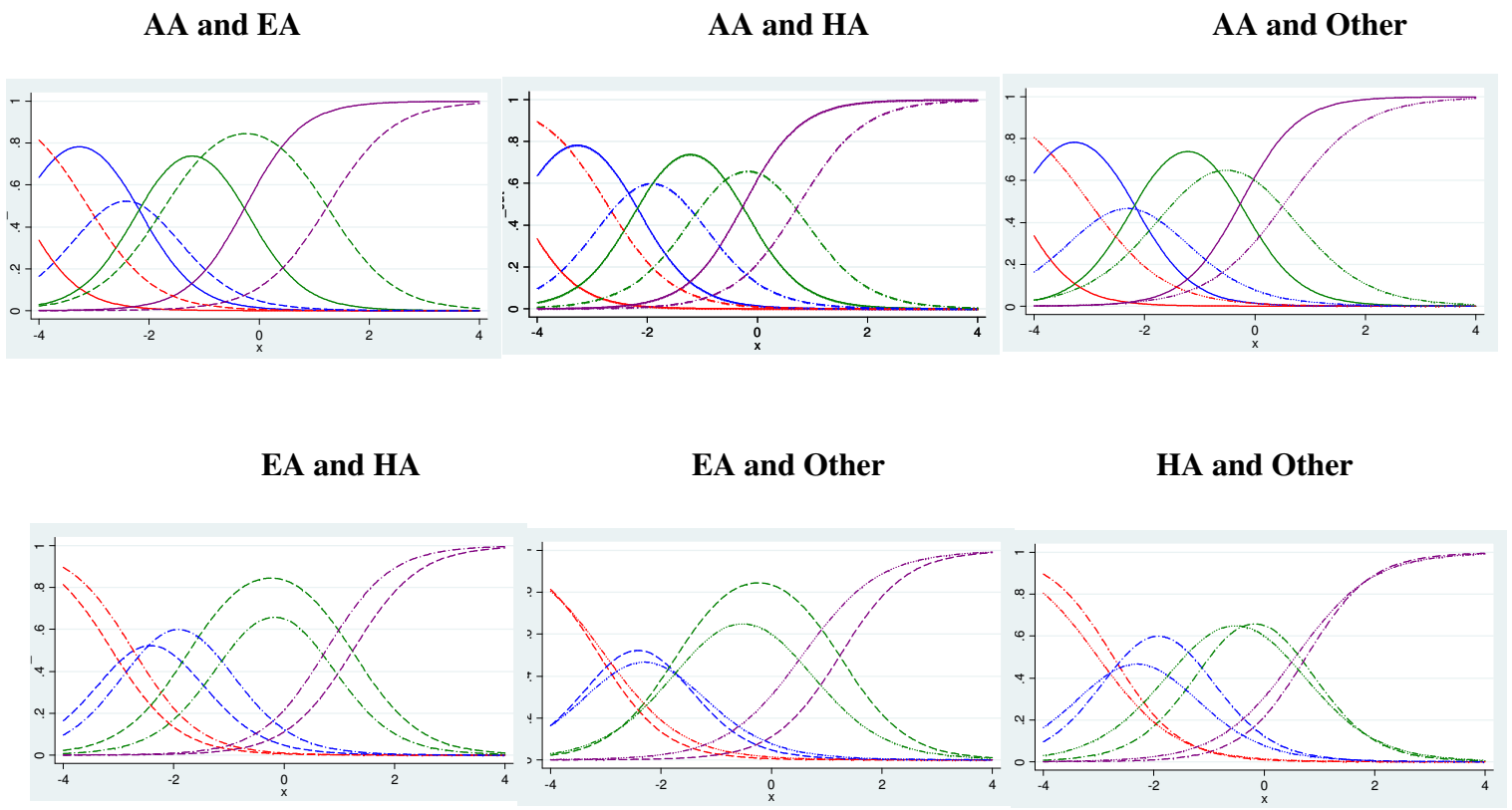


Figure **11. Item Characteristic Curves for Item 32 by Ethnicity**

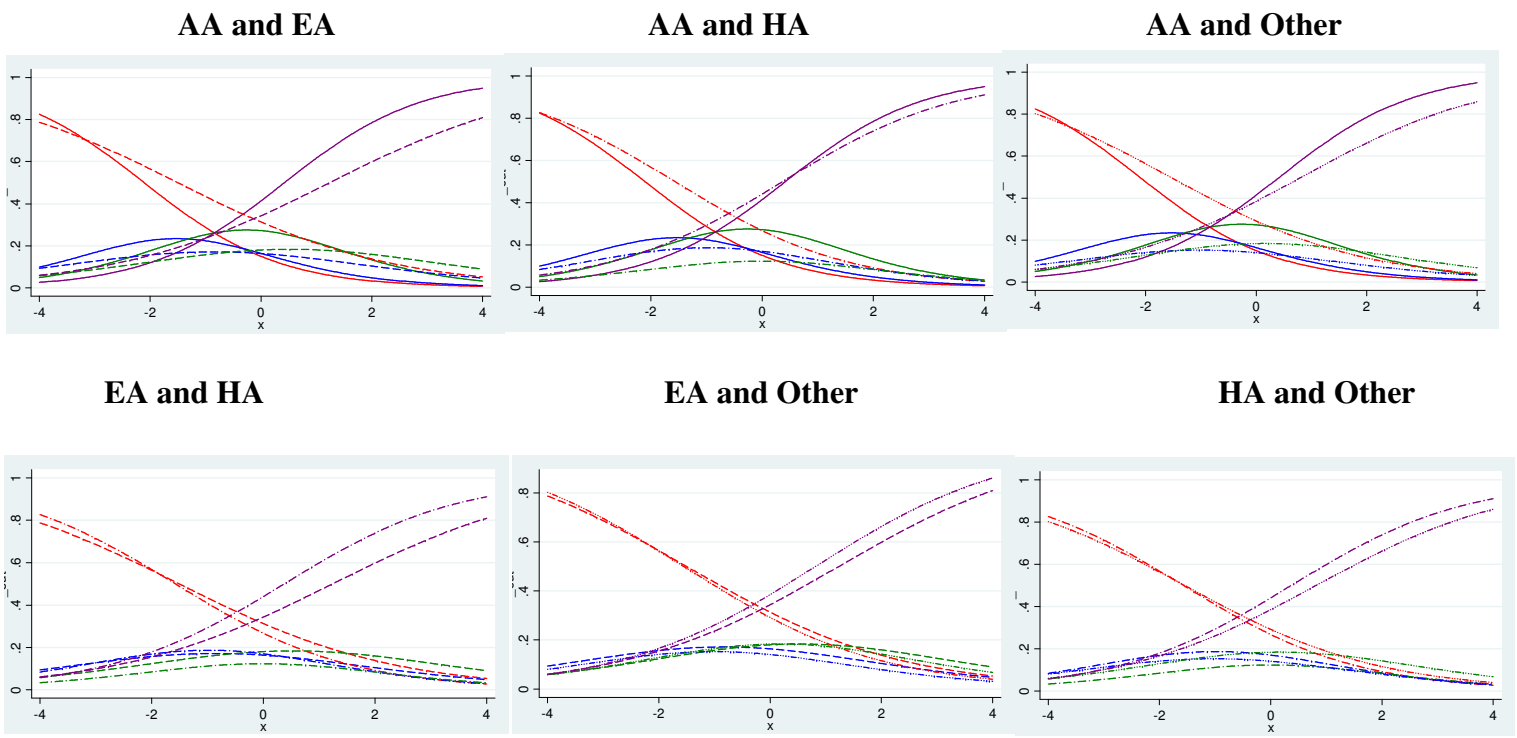


Figure **12. Item Characteristic Curves for Item 34 by Ethnicity**

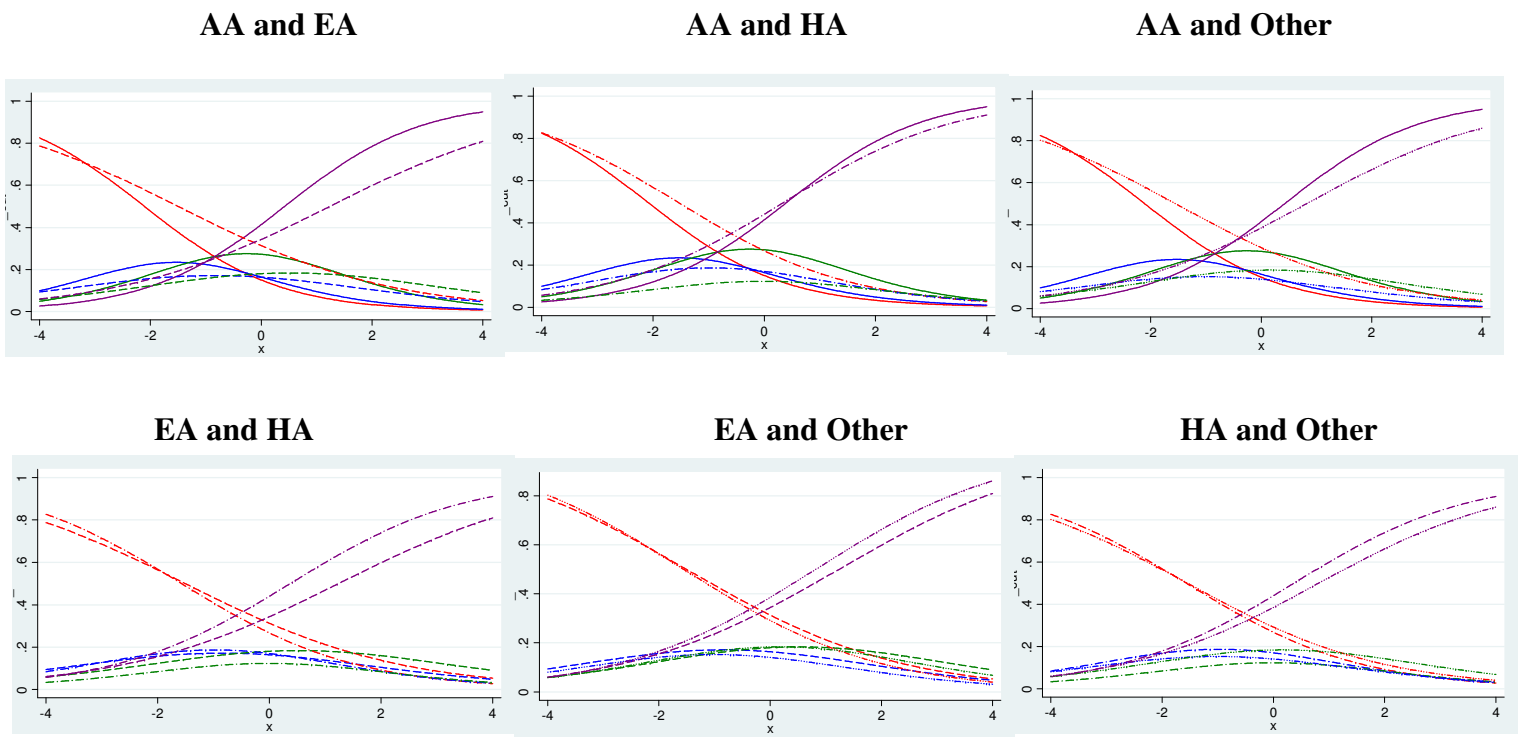


Figure **13. Item Characteristic Curves for Item 35 by Ethnicity**

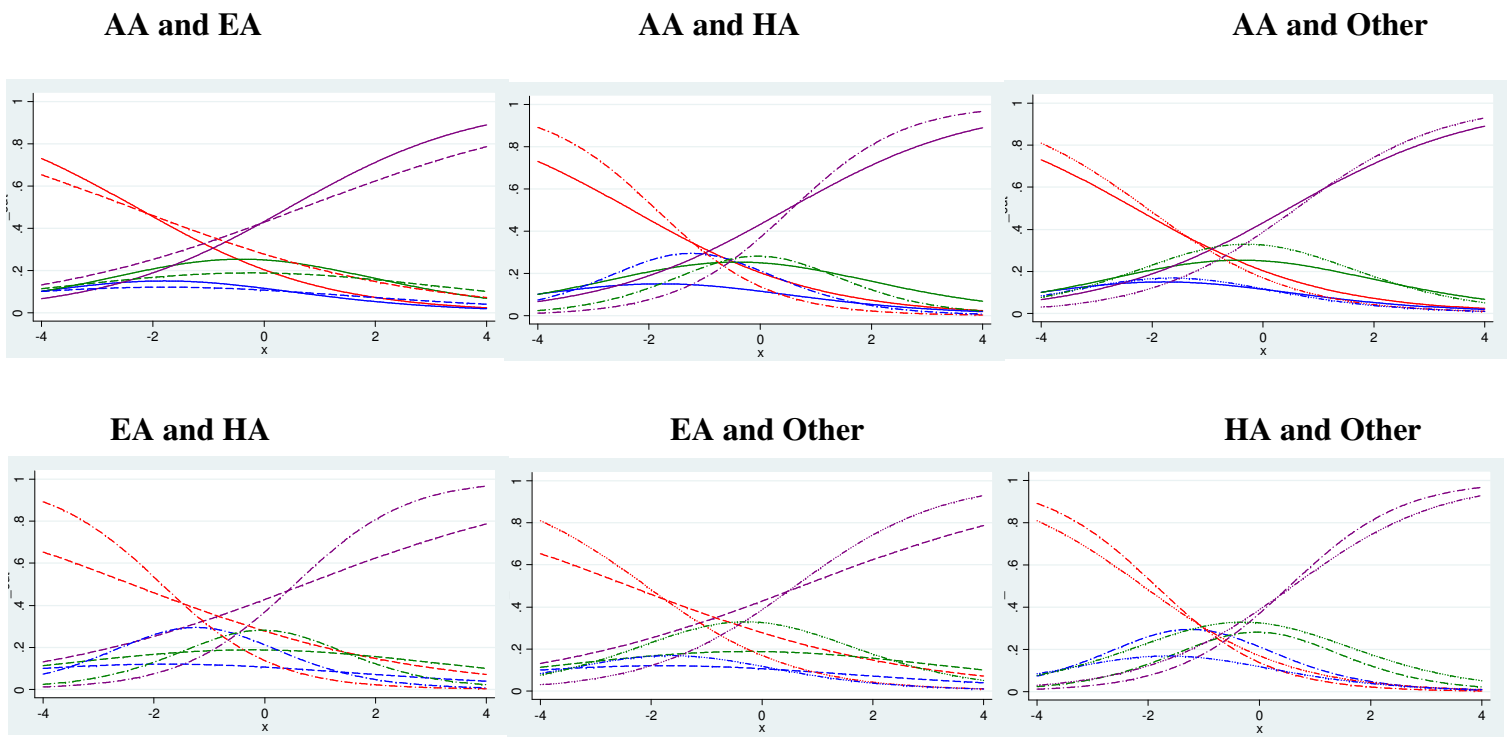


Figure **14. Item Characteristic Curves for Item 38 by Ethnicity**

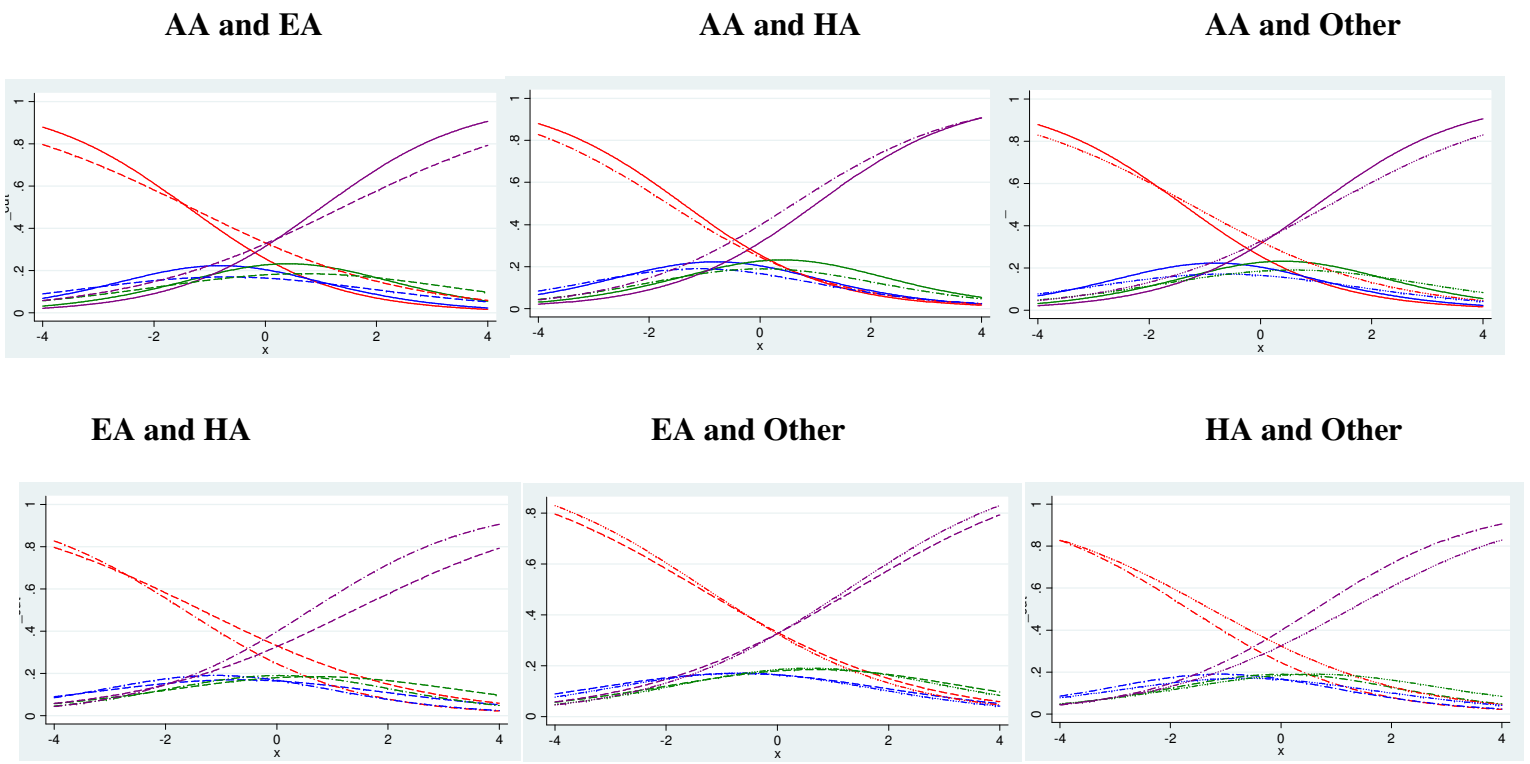


Figure **15. Item Characteristic Curves for Item 42 by Ethnicity**

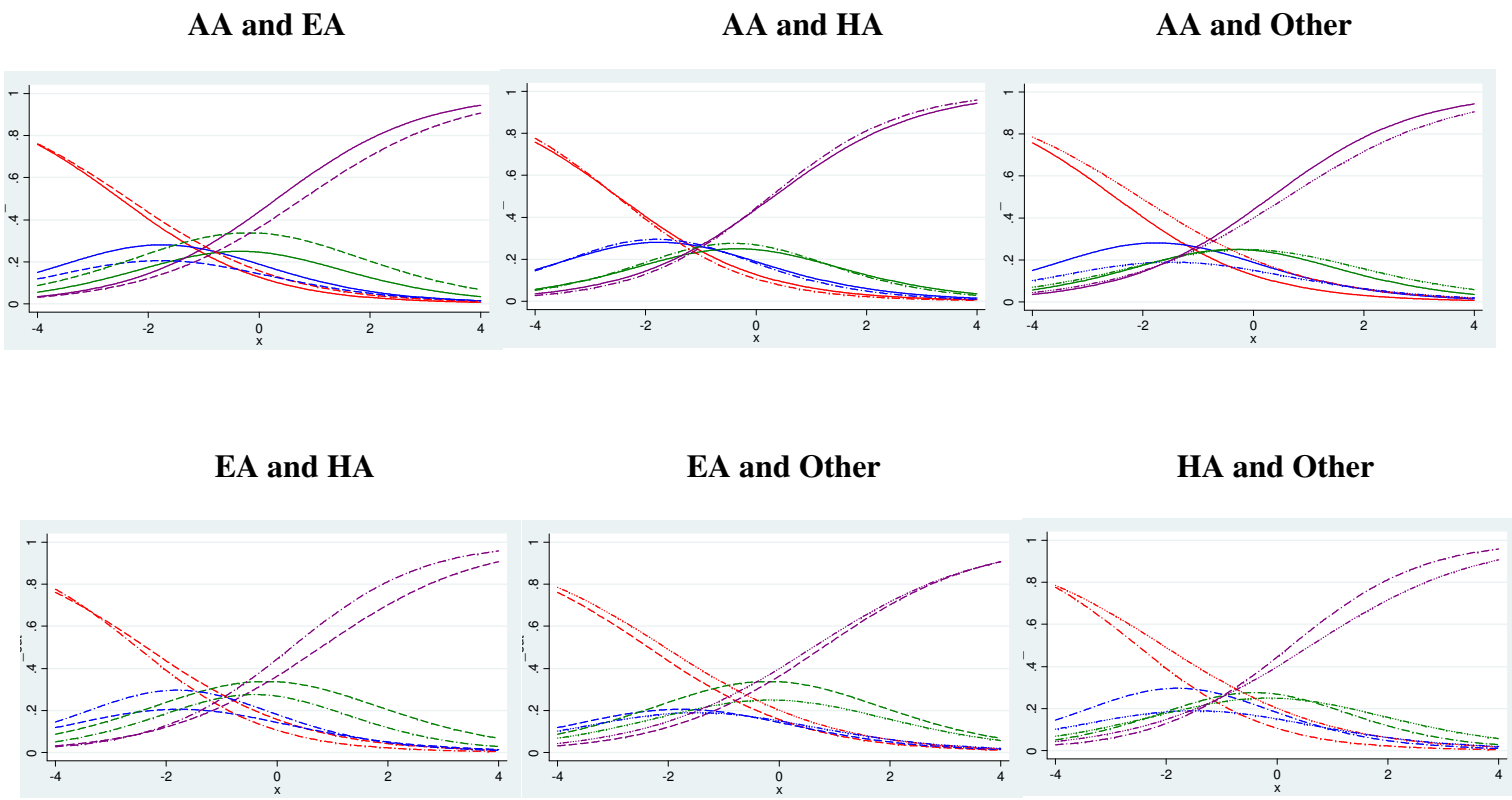


***Figure* 16. Item Characteristic Curves for Item 52 by Ethnicity**

Factor II: School Climate

The second factor was school climate which consisted of 5 items as shown in Table 8. Multilevel Item Response theory was applied to this factor, with the latent trait of interest being the teachers' perceptions of support provided by school climate to teachers in teaching African American students. Similar to the first factor, the analysis was first conducted to investigate the qualities of the items and how much they contribute to the school climate that provided a support system to help teachers with effective instruction to African American students. Then the effects of gender, ethnicity, and teaching experience of the teachers on their perceptions of school climate were estimated both for differential and the overall effect of these covariates on teachers' perceptions of school climate.

Table 8
Items in School Climate Factor

Question No.	Item
12	I feel supported by my building principal.
13	I feel I am supported by the administrative staff.
14	I feel supported by my professional colleagues
15	I believe I have opportunities to grow professionally as I fulfill duties at my ISD
17	I believe my contributions are appreciated by my colleagues.

The item location (threshold) and slope (discrimination) parameters for the items measuring teacher perceptions of school climate for various models are shown in Table 9. After considering the chi-square and the BIC values between the models, the 2-PL IRT model with varying thresholds was found to be the model of best fit for factor 2. The

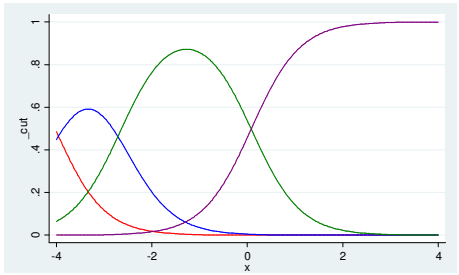
difference between the BICs of the 2-PL IRT model with varying thresholds and the 1-PL IRT model with varying thresholds was found to be 124.94 which show that the former model fits much better than the latter.

Items 12 and 13 had the lowest 1-2 threshold parameters, which show that teachers are less likely to strongly disagree with these items than the rest of the items in the factor. Items 13 and 17 had the highest 3-4 threshold parameters showing that teachers are less likely to strongly agree with these items compared to the other items. The discrimination parameters for the 5 items ranged from (2.283 to 0.661). Item 13 ("I feel I am supported by the administrative staff") had the highest discrimination parameter which means that this item contributed the most towards the latent trait, teacher perceptions about school climate in contributing to teaching African American students. Item 12 had a high discrimination parameter as well. Items 14, 15, and 17 had low discrimination parameters. The responses on these items tend to be more extreme than neutral and this can be seen from the item characteristic curves in Figure 17. Some of these values are also consistent with the factor pattern coefficients of the items on the factor in the previous study. Items 13, 12, and 14 had the highest factor pattern coefficients (0.765, 0.712, and 0.701 respectively) but only items 12 and 13 had high discrimination parameters. Item 14 had a low discrimination parameter and item 15 which had the lowest factor pattern coefficient had did not have the lowest discrimination parameter (0.829).

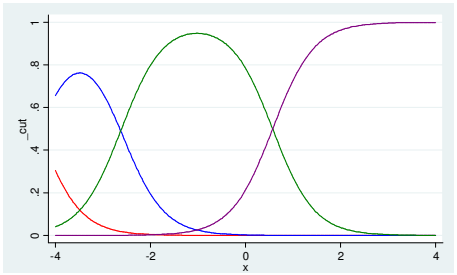
Table 9
Item Thresholds and Discrimination Parameters for Models 1 through 5-5 (Factor II)

Items	Parameters	1-PL (model 1)	2-PL (model 2)	1-PL (model 3)	2-PL (model 4)	gender cov. (model 5)	Ethnicity cov. (model 6)	Exp. cov. (model 7)
Item 12	Threshold 1-2	-2.980	-3.200	-2.766	-4.028	-3.796	-4.019	-4.298
	Threshold 2-3	-1.940	-2.090	-1.828	-2.642	-2.408	-2.633	-2.910
	Threshold 3-4	0.173	0.055	0.105	0.088	0.327	0.097	-0.175
	Discrimination	1.068	1.475	1.076	1.962	1.964	1.962	1.964
Item 13	Threshold 1-2	-3.237	-3.495	-2.689	-4.363	-4.086	-4.351	-4.657
	Threshold 2-3	-2.197	-2.386	-1.602	-2.610	-2.333	-2.598	-2.908
	Threshold 3-4	-0.084	-0.241	0.380	0.570	0.848	0.582	0.264
	Discrimination	1.068	1.440	1.076	2.283	2.281	2.282	2.275
Item 14	Threshold 1-2	-2.734	-3.148	-3.295	-2.798	-2.709	-2.790	-2.884
	Threshold 2-3	-1.694	-2.038	-2.398	-2.009	-1.925	-2.006	-2.101
	Threshold 3-4	0.419	0.107	-0.019	-0.007	0.078	-0.003	-0.099
	Discrimination	1.068	0.827	1.076	0.692	0.692	0.693	0.691
Item 15	Threshold 1-2	-3.254	-3.607	-2.571	-2.328	-2.226	-2.320	-2.438
	Threshold 2-3	-2.214	-2.497	-1.573	-1.412	-1.306	-1.404	-1.521
	Threshold 3-4	-0.101	-0.352	0.378	0.347	0.455	0.351	0.236
	Discrimination	1.068	1.093	1.076	0.829	0.830	0.832	0.833
Item 17	Threshold 1-2	-3.271	-3.674	-3.238	-2.698	-2.616	-2.693	-2.784
	Threshold 2-3	-2.231	-2.564	-1.844	-1.525	-1.444	-1.521	-1.613
	Threshold 3-4	-0.118	-0.419	0.600	0.502	0.582	0.506	0.414
	Discrimination	1.068	0.770	1.076	0.661	0.659	0.661	0.659
Model	Log Lkhd*	-4600.3	-4544.1	-4571.2	-4502.7	-4502.0	-4502.7	-4502.0
	BIC¥		100.40	14.92	124.94	-19.76(***N***)	-21.14(***N***)	-19.78(***N***)
	Random cov. Effect	-	--	--	--	0.198	0.002	-0.163
	P**	--	*0.00*	*0.00*	*0.00*	*0.24*(***N***)	*0.96*(***N***)	*0.24*(***N***)

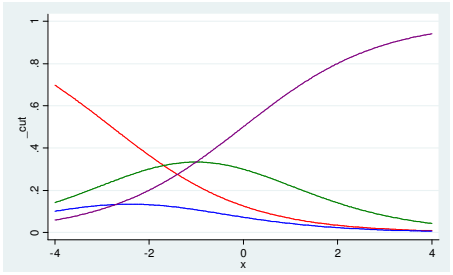
Note: * Log Lkhd represents the Log Likelihood of the model; BIC¥ represents the BIC of current model over previous model; ** P represents the probability that the previous model fits better than the current model



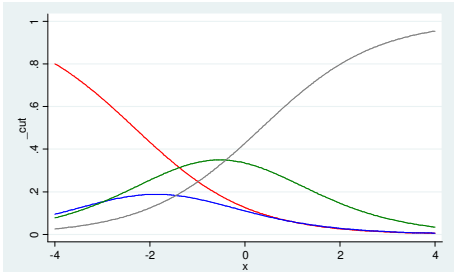
Item 12



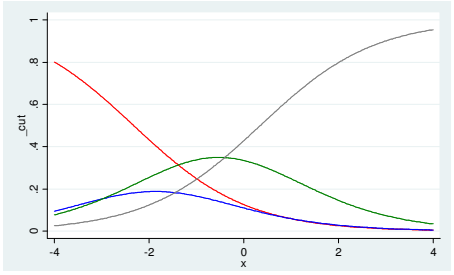
Item 13



Item 14



Item 15



Item 17

***Figure* 17. Item Characteristic Curves for Factor II – School Climate**

Table 9 further shows that when covariates gender, ethnicity, and teaching experience were included in the model there was no statistically significant improvement in the model fit as shown by the BIC and p-values. Therefore, it can be concluded that neither of the covariates had a statistically significant effect on school climate which means that the perceptions of teachers about the support provided by the school climate did not differ by their gender, ethnicity, or teaching experience.

Factor III: Culturally Responsive Management

The third factor was culturally responsive management which consisted of 3 items as shown in Table 10. As in the previous factors, Multilevel Item Response theory was applied to this factor, with the latent trait of interest being the extent of culturally responsive management exhibited by the teachers when teaching African American students. Similar to the previous factors, the analysis was first conducted to investigate the qualities of the items and how much they contribute to the extent of teachers' culturally responsive management while teaching African American students. Then the effects of gender, ethnicity, and teaching experience of the teachers on culturally responsive management were estimated both for differential and the overall effect of these covariates on the latent trait.

Unlike the previous factors, the model of best fit for this factor was a 1-PL model with varying threshold parameters (BIC = 24.55 less than the 1-PL model with fixed thresholds). The 2-PL model with varying threshold parameters did not converge and the 2-PL model with fixed parameters had a lower model fit than the 1-PL model with

varying thresholds as shown in Table 11. Therefore, the discrimination parameters were the same for all the items in this factor (2.039).

Table 10
Items in Culturally Responsive Management Factor

Question No.	Item
55	I believe I am able to effectively manage students from all racial groups.
56	I believe I have a clear understanding of the issues surrounding classroom management.
57	I believe I have clear understanding of the issues surrounding discipline.

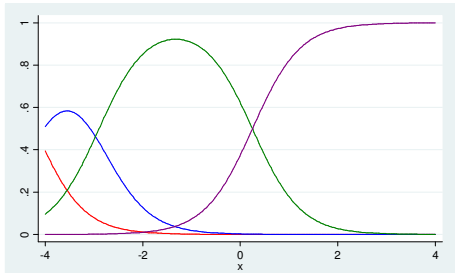
Items 56 and 57 had the lowest 1-2 threshold parameters, which show that teachers are less likely to strongly disagree with these items than the rest of the items in the factor. Item 56 also had the highest 3-4 threshold parameter showing that teachers are less likely to strongly agree with this item compared to the other items. The discrimination parameter was quite high for this factor. The item characteristic curves for the items forming this factor are shown in Figure 18. Items 56 and 57 had similar factor pattern coefficients (.911 and .903 respectively) in the previous study and have the same discrimination coefficients. However, item 55 which had a factor pattern coefficient of .784 had the same discrimination coefficient as well.

Table 11

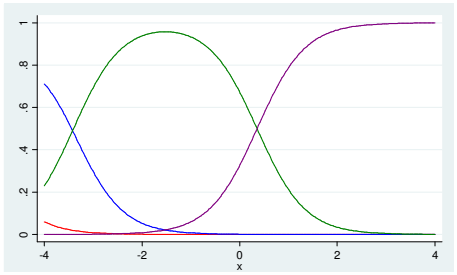
Item Thresholds and Discrimination Parameters for Models 1 through 5-5 (Factor III)

Items	Parameters	1-PL (model 1)	2-PL (model 2)	1-PL (model 3)	2-PL (model 4)	gender cov. (model 5)	Ethnicity cov. (model 6)	Exp. cov. (model 7)
Item 55	Threshold 1-2	-4.579	-4.458	-4.210	xx	-4.133	-5.006	-3.130
	Threshold 2-3	-3.070	-3.050	-2.900	xx	-2.825	-3.696	-1.820
	Threshold 3-4	0.386	0.404	0.250	xx	0.326	-0.533	1.320
	Discrimination	2.012	1.953	2.039	xx	2.039	2.000	2.012
Item 56	Threshold 1-2	-4.450	-4.393	-5.350	xx	-5.273	-6.123	-4.277
	Threshold 2-3	-2.941	-2.895	-3.405	xx	-3.333	-4.196	-2.327
	Threshold 3-4	0.515	0.559	0.357	xx	0.433	-0.427	1.428
	Discrimination	2.012	2.027	2.039	xx	2.039	2.001	2.012
Item 57	Threshold 1-2	-4.435	-4.365	-4.963	xx	-4.887	-5.740	-3.890
	Threshold 2-3	-2.926	-2.867	-3.356	xx	-3.281	-4.148	-2.280
	Threshold 3-4	0.530	0.586	0.249	xx	0.370	-0.490	1.365
	Discrimination	2.012	2.058	2.039	xx	2.039	2.001	2.012
Model	Log Lkhd*	-2184.9	-2184.4	-2169.6	***No convergence***	-2169.6	-2151.6	-2159.0
	BIC¥		-4.92	24.55§	--	-21.02 ***(N)***	14.98 (***Y***)	0.07 (***Y***)
	Random cov. Effect	--	--	--	--	0.063	-0.176	0.665
	P**	--	*0.57*	*0.00*	--	*0.72 (**N**)*	*0.00 (**Y**)*	*0.00 (**Y**)*

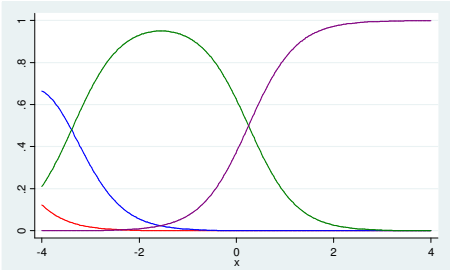
Note: * Log Lkhd represents the Log Likelihood of the model; BIC¥ represents the BIC of current model over previous model; ** P represents the probability that the previous model fits better than the current model; § represents the probability of model fit of model 3 over model 1



Item 55



Item 56



Item 57

***Figure* 18. Item Characteristic Curves for Factor III – Culturally Responsive Management**

Table 11 further shows that when covariates ethnicity and teaching experience were included in the model there was a statistically significant improvement in the model fit as shown by the BIC and p-values. Gender did not have a statistically significant impact on culturally responsive management. The estimates of overall and differential effects of the covariates on the individual items are shown in Table 12. Both teaching experience and ethnicity had a differential effect on items 55 and 57. Therefore, items 55 and 57 measured culturally responsive management of teachers from different ethnic groups differently. These items also measured culturally responsive management of novice and veteran teachers differently. This can also be seen from the item characteristic curves of the items by ethnicity and teaching experience in Figures 19-21.

The difference in the mean latent response for item, i between novice and veteran teachers can be given by the following formula:

*Difference in the mean latent response for item i = (random effect of the covariates * number of covariates) + (direct effect of the covariate on the item)*

Therefore, the difference in the mean latent responses for the item 55 between novice and veteran teachers was:

0.898 X 1 - 0.605 = 0.293

Table 12
Item Thresholds and Discrimination Parameters for Models with Statistically Significant Covariates (Factor III)

Covariate		Item 55	Item 56	Item 57
Ethn.	Threshold 1-2	-5.490	-5.974	-5.512
	Threshold 2-3	-4.158	-4.024	-3.190
	Threshold 3-4	-0.899	-0.271	-0.266
	Discrimination	2.027	2.005	2.009
Model	Log Lkhd*	-2139.640	-2149.510	-2147.290
	BIC¥	17.915	-1.825	2.605
	Random cov. Effect	-0.135	-0.194	-0.202
	Cov. Effect on Item	-0.125	0.053	0.076
	P**	0.000	0.040	0.003
	Presence of **DIF**	**YES**	**NO**	**YES**
Exp.	Threshold 1-2	-3.750	-4.090	-3.479
	Threshold 2-3	-2.430	-2.132	-1.849
	Threshold 3-4	0.728	1.650	1.885
	Discrimination	2.044	2.015	2.029
Model	Log Lkhd*	-2146.520	-2157.800	-2152.510
	BIC¥	19.035	-3.515	7.055
	Random cov. Effect	0.898	0.604	0.527
	Cov. Effect on Item	-0.605	0.053	0.454
	P**	0.000	0.113	0.000
	Presence of **DIF**	**YES**	**NO**	**YES**

Note: * Log Lkhd represents the Log Likelihood of the model; BIC¥ represents the BIC of current model over previous model; ** P represents the probability that the previous model fits better than the current model

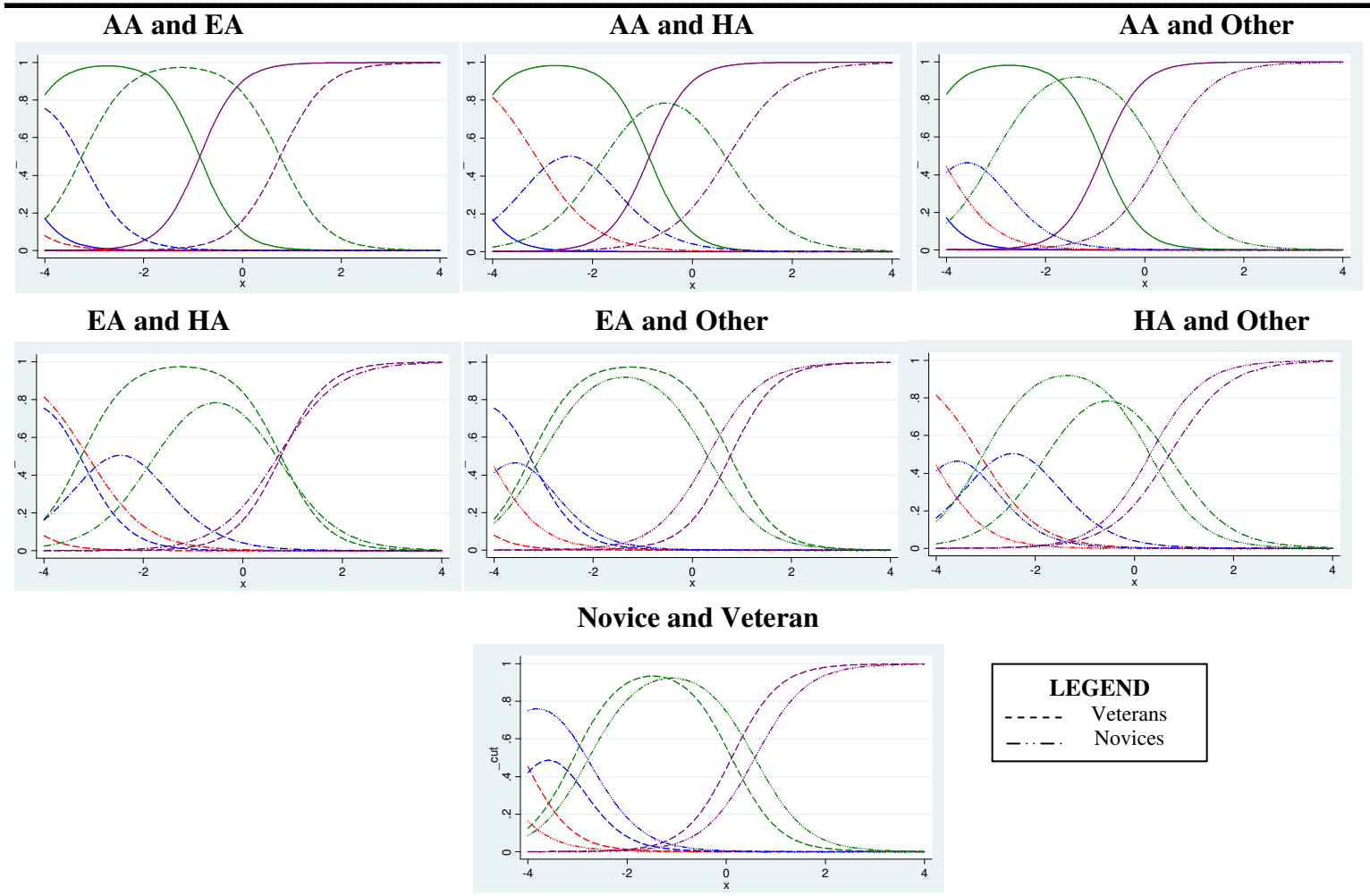


Figure **19. Item Characteristic Curves for Item 55 by Ethnicity and Teaching Experience**

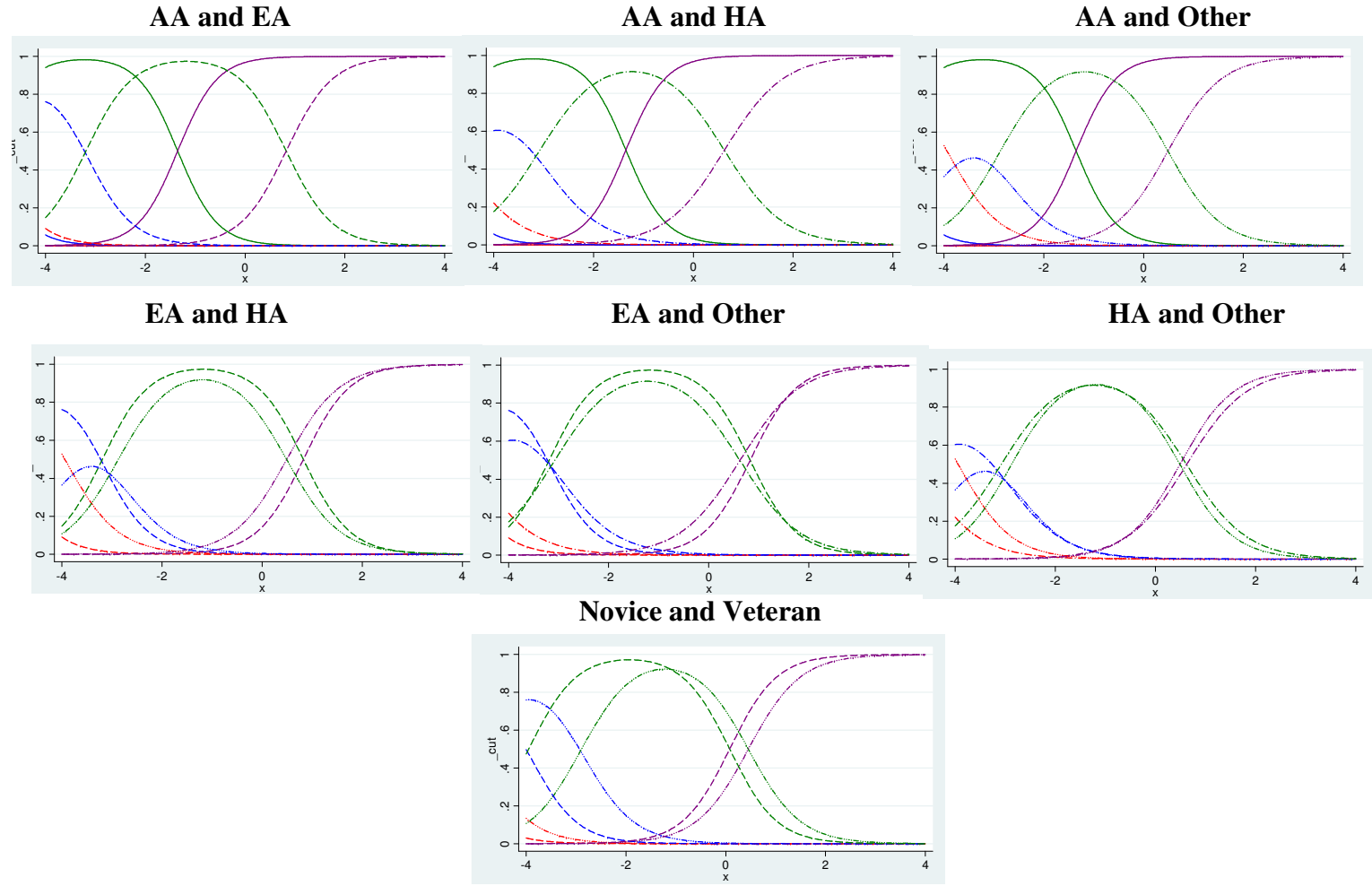


Figure **20. Item Characteristic Curves for Item 56 by Ethnicity and Teaching Experience**

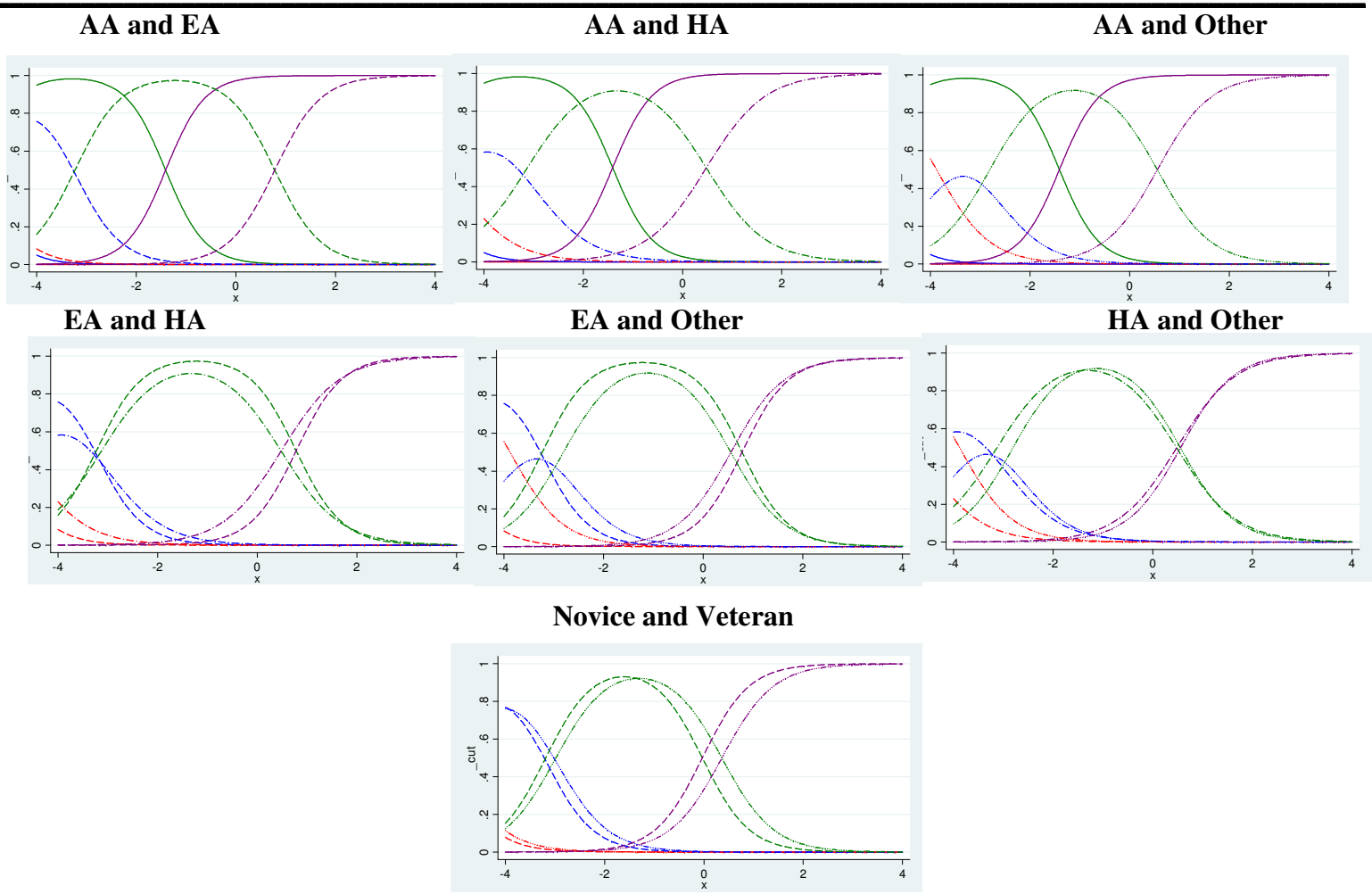


Figure 21. Item Characteristic Curves for Item 57 by Ethnicity and Teaching Experience

Factor IV: Home and Community Support

The fourth factor, home and community support consisted of 4 items as shown in Table 13. The item location (threshold) and slope (discrimination) parameters for the items measuring teacher perceptions of home and community support for various models are shown in Table 14. After considering the chi-square and the BIC values between the models, the 2-PL IRT model with varying thresholds was found to be the model of best fit for Factor IV. The difference between the BICs of the 2-PL IRT model with varying thresholds and the 1-PL IRT model with varying thresholds was found to be 141.25 which show that the former model fits much better than the latter.

Table 13
Items in Home and Community Support Factor

Question No.	Item
19	I believe "all" students in my ISD are treated equitably regardless of race, culture, disability, gender or social economic status.
20	I believe my ISD families are supportive of our mission to effectively teach all students.
21	I believe my ISD families of African American students are supportive of our mission to effectively teach all students.
22	I believe the district has strong support for academic excellence from our surrounding community (civic, church, business).

Items 20 and 21 had the lowest 1-2 threshold parameters, which show that teachers are less likely to strongly disagree with these items than the rest of the items in the factor. Items 20 and 21 also had the highest 3-4 threshold parameters showing that teachers are less likely to strongly agree with these items as well when compared to the other items. The discrimination parameters for the 4 items ranged from (1.716 to 0.570). Item 20 ("I believe my ISD families are supportive of our mission to effectively teach all students") had the highest discrimination parameter which means that this item contributed the most towards the latent trait, teacher perceptions about home and community support for teaching African American students. Item 21 had a high discrimination parameter as well. Items 19 and 22 had the lowest discrimination parameters. Item 19 almost dichotomizes the responses while the responses on item 22 had very few responses in the middle categories and most responses in the extremes. This can be seen from the item characteristic curves in Figure 22. All these values are consistent with the factor pattern coefficients of the items on the factor in the previous study. Items 19, 20, 21, and 22 had factor pattern coefficient values of .480, .775, .804, and .581 respectively.

Table 14

Item Thresholds and Discrimination Parameters for Models 1 through 5-5 (Factor IV)

Items	Parameters	1-PL (model 1)	2-PL (model 2)	1-PL (model 3)	2-PL (model 4)	gender cov. (model 5)	Ethnicity cov. (model 6)	Exp. cov. (model 7)
Item 19	Threshold 1-2	-2.669	-2.703	-2.145	-1.822	-1.770	-1.766	-2.013
	Threshold 2-3	-1.158	-1.197	-0.954	-0.803	-0.746	-0.746	-0.994
	Threshold 3-4	0.785	0.759	0.504	0.444	0.504	0.504	0.254
	Discrimination	0.961	1.153	1.005	0.570	0.573	0.573	0.569
Item 20	Threshold 1-2	-3.085	-3.149	-2.470	-3.381	-3.233	-3.216	-3.946
	Threshold 2-3	-1.575	-1.643	-0.821	-1.125	-0.970	-0.953	-1.692
	Threshold 3-4	0.368	0.313	1.361	1.837	2.002	2.023	1.266
	Discrimination	0.961	0.997	1.005	1.716	1.723	1.723	1.701
Item 21	Threshold 1-2	-3.288	-3.357	-2.215	-2.855	-2.702	-2.676	-3.383
	Threshold 2-3	-1.778	-1.851	-0.646	-0.838	-0.690	-0.672	-1.366
	Threshold 3-4	0.165	0.105	1.645	2.140	2.282	2.289	1.612
	Discrimination	0.961	0.945	1.005	1.593	1.586	1.577	1.582
Item 22	Threshold 1-2	-3.228	-3.306	-2.329	-2.135	-2.060	-2.050	-2.402
	Threshold 2-3	-1.717	-1.800	-0.632	-0.575	-0.501	-0.492	-0.841
	Threshold 3-4	0.226	0.156	1.473	1.348	1.422	1.429	1.081
	Discrimination	0.961	0.808	1.005	0.805	0.803	0.801	0.800
Model	Log Lkhd*	-4334.0	-4323.1	-4261.8	-4186.7	-4186.3	-4185.4	-4182.6
	BIC¥		12.86	113.44	141.25	-20.36(***N***)	-18.59(***N***)	-12.93(***N***)
	Random cov. Effect	--	--	--	--	0.043	0.013	-0.117
	P**	--	*0.00*	*0.00*	*0.00*	*0.38*(***N***)	*0.11*(***N***)	*0.00*(***Y***)

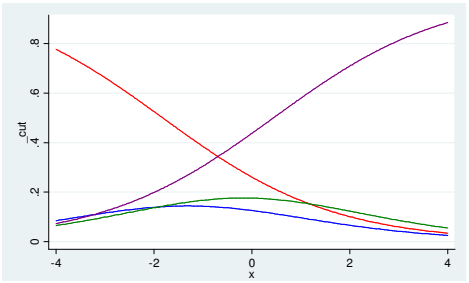
Note: * Log Lkhd represents the Log Likelihood of the model; BIC¥ represents the BIC of current model over previous model; ** P represents the probability that the previous model fits better than the current model

Table 14 further shows that when teaching experience was included in the model there was a statistically significant improvement in the model fit as shown by the BIC and p-values. Gender and ethnicity of the teacher did not have a statistically significant impact on the home and community support received by the teachers. The estimates of overall and differential effects of the covariates on the latent trait and the individual items are shown in Table 15. However, when the effect of teaching experience on the individual items 19, 20, 21, and 22 were computed, there was no differential effect of teaching experience on the home and community support received by the teachers. This can also be seen from the item characteristic curves of the items by teaching experience in Figure 23. Therefore, there was an overall effect of teaching experience on the factor but there was no individual or differential effect of teaching experience on each of the items.

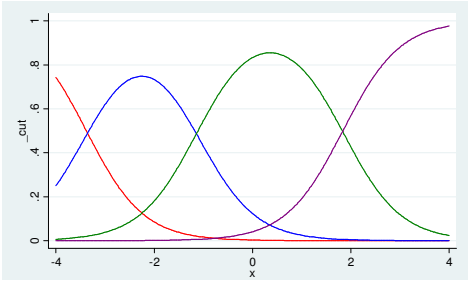
Table 15
Item Thresholds and Discrimination Parameters for Models with Statistically Significant Covariates (Factor IV)

Covariate		Item 19	Item 20	Item 21	Item 22
Exp.	Threshold 1-2	-2.135	-3.924	-3.383	-2.362
	Threshold 2-3	-1.110	-1.665	-1.364	-0.801
	Threshold 3-4	0.133	1.299	1.614	1.123
	Discrimination	0.564	1.710	1.636	0.803
Model	Log Lkhd*	-4182.010	-4182.580	-4182.620	-4182.550
	BIC¥	-19.929	-21.059	-21.139	-20.996
	Random cov. Effect	-0.111	-0.122	-0.117	-0.119
	Cov. Effect on Item	-0.082	0.033	0.002	0.029
	P**	0.271	0.775	0.982	0.780
	Presence of **DIF**	**NO**	**NO**	**NO**	**NO**

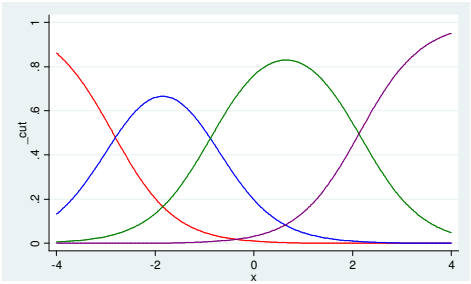
Note: * Log Lkhd represents the Log Likelihood of the model; BIC¥ represents the BIC of current model over previous model; ** P represents the probability that the previous model fits better than the current model



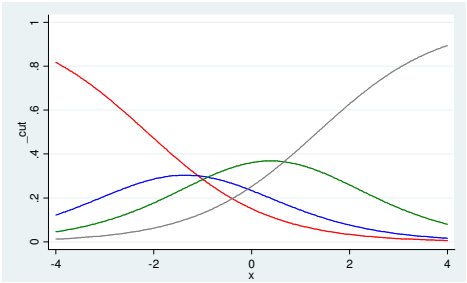
Item 19



Item 20

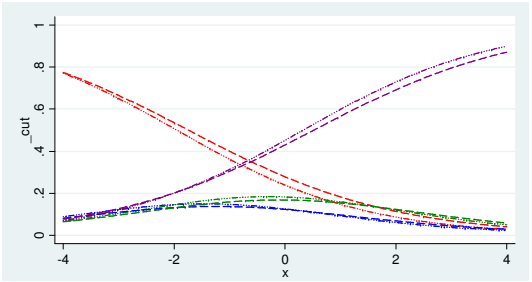


Item 21

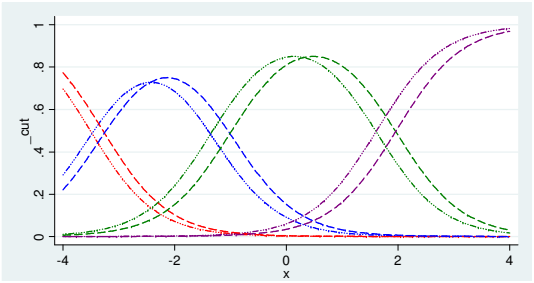


Item 22

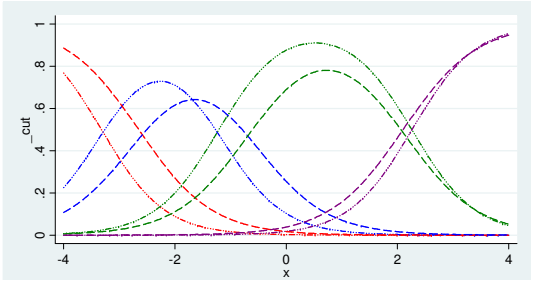
***Figure* 22. Item Characteristic Curves for Factor IV – Home and Community Support**



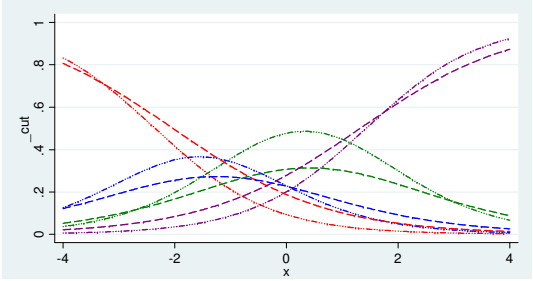
Item 19



Item 20



Item 21



Item 22

***Figure* 23. Item Characteristic Curves for Factor IV by Teaching Experience**

Factor V: Cultural Sensitivity

The fifth factor, cultural sensitivity consisted of 5 items as shown in Table 16. The item location (threshold) and slope (discrimination) parameters for the items measuring the cultural sensitivity of teachers for various models are shown in Table 17. After considering the chi-square and the BIC values between the models, the 2-PL IRT model with varying thresholds was found to be the model of best fit for factor 5. The difference between the BICs of the 2-PL IRT model with varying thresholds and the 1-PL IRT model with varying thresholds was found to be 68.52 which show that the former model fits much better than the latter.

Table 16
Items in Cultural Sensitivity Factor

Question No.	Item
37	I believe it is important to identify with the racial groups of the students I serve.
39	I believe I am comfortable with people who exhibit values or beliefs different from my own.
40	I believe cultural views of a diverse community should be included in the school's yearly program planning.
41	I believe it is necessary to include on-going family input in program planning.
50	I believe Individualized Education Program meetings or planning should be scheduled for the convenience of the family.

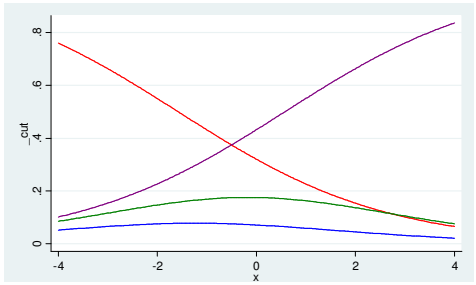
Table 17

Item Thresholds and Discrimination Parameters for Models 1 through 5-5 (Factor V)

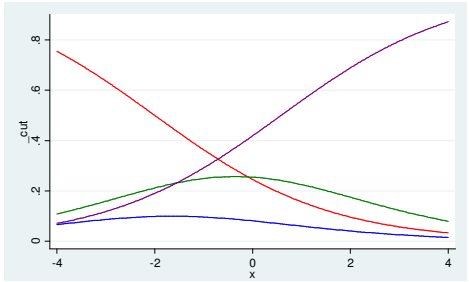
Items	Parameters	1-PL (model 1)	2-PL (model 2)	1-PL (model 3)	2-PL (model 4)	gender cov. (model 5)	Ethnicity cov. (model 6)	Exp. cov. (model 7)
Item 37	Threshold 1-2	-1.965	-1.976	-1.670	-1.577	-1.663	-1.635	-1.670
	Threshold 2-3	-1.060	-1.078	-0.978	-0.919	-1.004	-0.976	-1.012
	Threshold 3-4	0.737	0.729	0.606	0.574	0.490	0.520	0.480
	Discrimination	0.637	0.746	0.643	0.475	0.476	0.478	0.474
Item 39	Threshold 1-2	-2.494	-1.818	-2.067	-1.999	-2.101	-2.063	-2.110
	Threshold 2-3	-0.938	-0.920	-1.327	-1.289	-1.389	-1.353	-1.399
	Threshold 3-4	0.960	0.887	0.601	0.584	0.485	0.520	0.474
	Discrimination	0.637	0.561	0.643	0.560	0.560	0.558	0.559
Item 40	Threshold 1-2	-2.208	-1.494	-2.363	-3.123	-3.330	-3.265	-3.385
	Threshold 2-3	-0.698	-0.596	-1.636	-2.177	-2.389	-2.326	-2.432
	Threshold 3-4	1.246	1.211	0.312	0.395	0.166	0.245	0.141
	Discrimination	0.637	0.812	0.643	1.292	1.276	1.290	1.293
Item 41	Threshold 1-2	-2.421	-1.735	-2.564	-2.832	-2.998	-2.934	-3.000
	Threshold 2-3	-0.911	-0.837	-1.371	-1.517	-1.680	-1.620	-1.689
	Threshold 3-4	1.033	0.971	0.552	0.604	0.447	0.502	0.430
	Discrimination	0.637	0.670	0.643	0.888	0.893	0.888	0.885
Item 50	Threshold 1-2	-3.056	-2.383	-1.753	-1.622	-1.691	-1.665	-1.697
	Threshold 2-3	-1.546	-1.485	-0.591	-0.540	-0.609	-0.584	-0.615
	Threshold 3-4	0.398	0.323	1.095	1.004	0.936	0.960	0.928
	Discrimination	0.637	0.559	0.601	0.473	0.393	0.416	0.383
Model	Log Lkhd*	-5312.5	-5298.4	-5273.3	-5234.5	-5233.3	-5233.2	-5233.3
	BIC¥		19.20	41.04	68.52	-18.72(***N***)	-18.45(***N***)	-18.74(***N***)
	Random cov. Effect	--	--	--	--	-0.07	-0.012	-0.057
	P**	--	*0.00*	*0.00*	*0.00*	*0.12*(***N***)	*0.10*(***N***)	*0.12*(***N***)

Note: * Log Lkhd represents the Log Likelihood of the model; BIC¥ represents the BIC of current model over previous model; ** P represents the probability that the previous model fits better than the current model

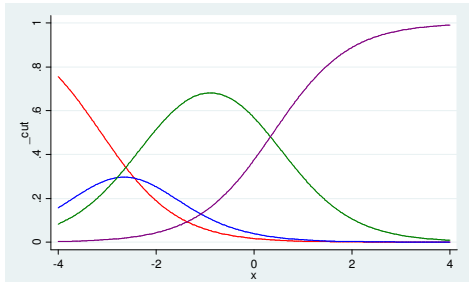
Items 40 and 41 had the lowest 1-2 threshold parameters, which show that teachers are less likely to strongly disagree with these items than the rest of the items in the factor. Item 41 also had the highest 3-4 threshold parameter showing that teachers are less likely to strongly agree with this item as well when compared to the other items. The discrimination parameters for the 5 items ranged from (1.292 to 0.473). Item 40 ("I believe cultural views of a diverse community should be included in the school's yearly program planning") had the highest discrimination parameter, which means that this item contributed the most towards the latent trait, cultural sensitivity of teachers. Items 37, 39, and 50 had the lowest discrimination parameters. In fact, all these 3 items almost dichotomize the responses, as can be seen from the item characteristic curves in Figure 24. When compared to factor pattern coefficients of these items, item 39 had a high factor pattern coefficient of 0.612 but a very low discrimination parameter. Table 17 further shows that when covariates such as gender, ethnicity, and teaching experience was included in the model there was no statistically significant improvement in the model fit as shown by the BIC and p-values. Therefore, gender, ethnicity, and teaching experience did not have a statistically significant impact on the cultural sensitivity of teachers.



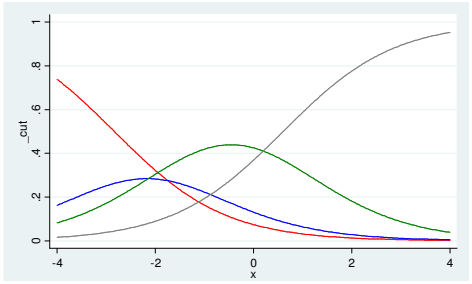
Item 37



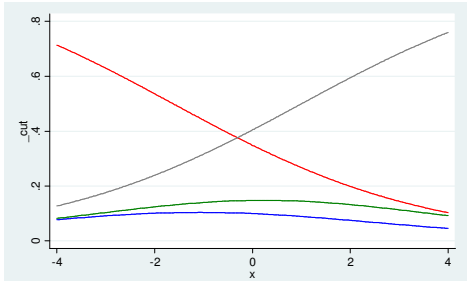
Item 39



Item 40



Item 41



Item 50

***Figure* 24. Item Characteristic Curves for Factor V – Cultural Sensitivity**

Factor VI: Curriculum and Instructional Strategies

The sixth factor, curriculum and instructional strategies consisted of 4 items as shown in Table 18. As in the previous factors, Multilevel Item Response theory was applied to this factor, with the latent trait of interest being the curriculum and instructional strategies of teachers when teaching African American students. Similar to the previous factors, the analysis was first conducted to investigate the qualities of the items and how much they contribute to the curriculum and instructional strategies of teachers that helps them teach African American students effectively. Then the effects of gender, ethnicity, and teaching experience of the teachers on the latent trait were estimated.

Table 18
Items in Curriculum and Instructional Strategies Factor

Question No.	Item
26	I believe the in-service training this past year assisted me in improving my teaching strategies.
27	I believe I am culturally responsive in my teaching behaviors.
28	I believe cooperative learning is an integral part of my ISD teaching and learning philosophy
51	I believe frequently used material within my class represents at least three different ethnic groups.

The item location (threshold) and slope (discrimination) parameters for the items measuring the curriculum and instructional strategies of teachers for various models are shown in Table 19. After considering the chi-square and the BIC values between the models, the 2-PL IRT model with varying thresholds was found to be the model of best fit for factor 6. The difference between the BICs of the 2-PL IRT model with varying thresholds and the 1-PL IRT model with varying thresholds was found to be 52.07 which show that the former model fits much better than the latter.

Items 27 and 28 had the lowest 1-2 threshold parameters, which show that teachers are less likely to strongly disagree with these items than the rest of the items in the factor. Item 27 and 28 also had the highest 3-4 threshold parameter showing that teachers are less likely to strongly agree with these items as well when compared to the other items. The discrimination parameters for the 4 items ranged from (1.202 to 0.343). Item 28 (I believe cooperative learning is an integral part of my ISD teaching and learning philosophy) had the highest discrimination parameter which means that this item contributed the most towards the latent trait, curriculum and instructional strategies of teachers. Items 26 and 51 had very low discrimination parameters. In fact these 2 items almost dichotomize the responses as can be seen from the item characteristic curves in Figure 25. These two items had the lowest factor pattern coefficients as well (.5 and .423 respectively).

Table 19

Item Thresholds and Discrimination Parameters for Models 1 through 5-5 (Factor VI)

Items	Parameters	1-PL (model 1)	2-PL (model 2)	1-PL (model 3)	2-PL (model 4)	gender cov. (model 5)	Ethnicity cov. (model 6)	Exp. cov. (model 7)
Item 26	Threshold 1-2	-1.625	-1.630	-1.516	-1.439	-1.353	-1.458	-1.630
	Threshold 2-3	-0.497	-0.501	-0.450	-0.425	-0.340	-0.445	-0.615
	Threshold 3-4	1.478	1.470	1.332	1.262	1.349	1.242	1.082
	Discrimination	0.629	0.715	0.631	0.490	0.490	0.488	0.501
Item 27	Threshold 1-2	-0.460	-0.471	-2.943	-3.248	-3.123	-3.277	-3.495
	Threshold 2-3	0.667	0.658	-1.960	-2.165	-2.038	-2.201	-2.435
	Threshold 3-4	2.643	2.629	0.412	0.449	0.613	0.413	0.132
	Discrimination	0.629	0.587	0.631	0.875	0.900	0.872	0.831
Item 28	Threshold 1-2	-0.435	-0.436	-2.956	-3.809	-3.516	-3.878	-4.390
	Threshold 2-3	0.692	0.693	-1.869	-2.423	-2.160	-2.485	-2.965
	Threshold 3-4	2.668	2.664	0.352	0.444	0.637	0.398	-0.008
	Discrimination	0.629	0.675	0.631	1.202	1.152	1.212	1.255
Item 51	Threshold 1-2	-1.511	-1.524	-1.781	-1.623	-1.563	-1.636	-1.737
	Threshold 2-3	-0.384	-0.395	-0.487	-0.441	-0.382	-0.455	-0.563
	Threshold 3-4	1.592	1.576	1.207	1.095	1.157	1.082	0.969
	Discrimination	0.629	0.549	0.631	0.343	0.345	0.341	0.327
Model	Log Lkhd*	-4072.2	-4069.4	-4036.8	-4006.2	-4005.2	-4006.1	-4002.4
	BIC¥		-3.52	61.84§	52.07	-19.07(***N***)	-20.83(***N***)	-13.55(***N***)
	Random cov. Effect	--	--	--	--	0.07	-0.004	-0.115
	P**	--	*0.13*	*0.00*	*0.00*	*0.15*(***N***)	*0.58*(***N***)	*0.00*(***Y***)

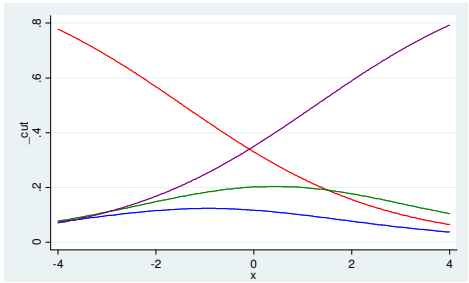
Note: * Log Lkhd represents the Log Likelihood of the model; BIC¥ represents the BIC of current model over previous model; ** P represents the probability that the previous model fits better than the current model; § represents the probability of model fit of model 3 over model 1

Table 19 further shows that when teaching experience was included in the model there was a statistically significant improvement in the model fit as shown by the BIC and p-values. Gender and ethnicity of the teacher did not have a statistically significant impact on the curriculum and instructional strategies of the teachers. The estimates of overall and differential effects of the covariates on the latent trait and the individual items are shown in Table 20. When the effect of teaching experience on the individual items 26, 27, 28, and 51 were estimated, there was a differential effect of teaching experience on item 26 only. The item characteristic curves of the items by teaching experience are shown in Figure 26.

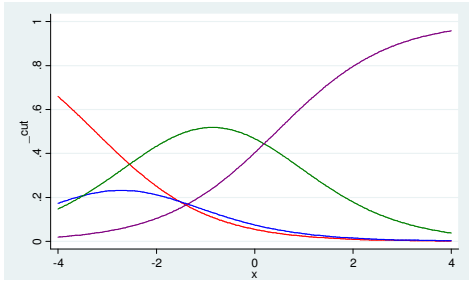
Table 20
Item Thresholds and Discrimination Parameters for Models with Statistically Significant Covariates (Factor VI)

Covariate		Item 26	Item 27	Item 28	Item 51
Exp.	Threshold 1-2	-2.063	-3.227	-4.390	-1.423
	Threshold 2-3	-1.040	-2.194	-3.006	-0.238
	Threshold 3-4	0.666	0.447	0.116	1.307
	Discrimination	0.483	0.897	1.201	0.354
Model	Log Lkhd*	-3994.630	-3998.400	-4001.600	-2997.930
	BIC¥	6.530	-0.987	-7.397	-0.057
	Random cov. Effect	-0.070	-0.168	-0.074	-0.137
	Cov. Effect on Item	-0.306	0.294	-0.171	-0.026
	P**	0.000	0.000	0.198	0.003
	Presence of **DIF**	**YES**	**NO**	**NO**	**NO**

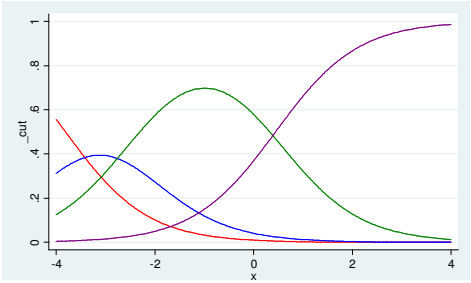
Note: * Log Lkhd represents the Log Likelihood of the model; BIC¥ represents the BIC of current model over previous model; ** P represents the probability that the previous model fits better than the current model



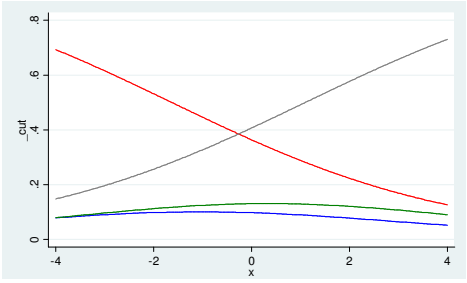
Item 26



Item 27

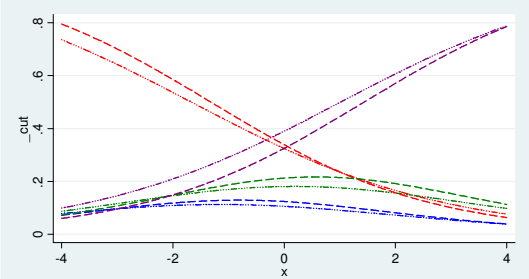


Item 28

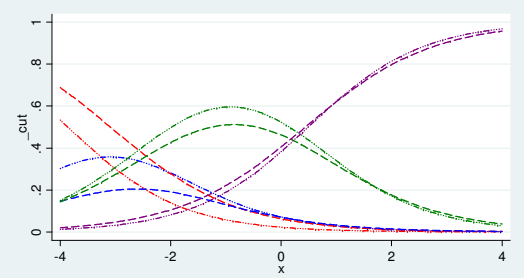


Item 51

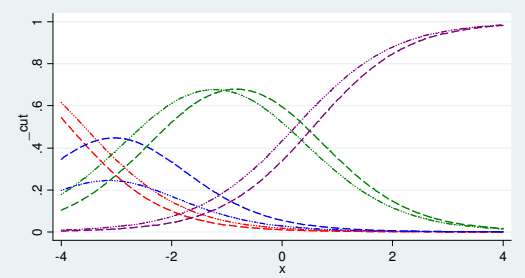
***Figure* 25. Item Characteristic Curves for Factor VI – Curriculum and Instructional Strategies**



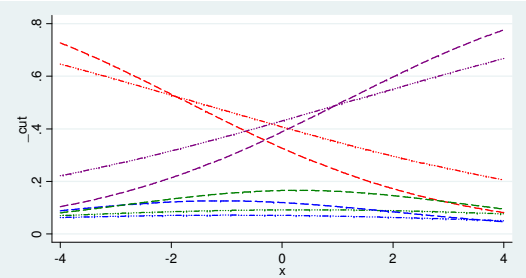
Item 26



Item 27



Item 28



Item 51

***Figure* 26. Item Characteristic Curves for Factor VI by Teaching Experience**

Factor VII: Cultural Awareness

The seventh factor, cultural awareness consisted of 3 items as shown in Table 21. As in the previous factors, Multilevel Item Response theory was applied to this factor, with the latent trait of interest being the level of cultural awareness of teachers. Similar to the previous factors, the analysis was first conducted to investigate the qualities of the items and how much they contribute to the cultural awareness of teachers that helps them teach African American students effectively. Then the effects of gender, ethnicity, and teaching experience of the teachers on the latent trait were estimated.

Table 21
Items in Cultural Awareness Factor

Question No.	Item
46	I believe that in a society with as many racial groups as the United States, I would accept the use of ethnic jokes or phrases by students.
47	I believe there are times when “racial statements” should be ignored.
48	I believe a child should be referred “for testing” if learning difficulties appear to be due to cultural differences.

The item location (threshold) and slope (discrimination) parameters for the items measuring the cultural awareness teachers for various models are shown in Table 22. After considering the chi-square and the BIC values between the models, the 2-PL IRT model with fixed thresholds was found to be the model of best fit for factor 7. The difference between the BICs of the 2-PL IRT model with fixed thresholds and the 1-PL IRT model with fixed thresholds was found to be 40.82. Subsequent models such as 1-PL model with varying thresholds and 2-PL model with varying thresholds did not improve the fit of the model. Therefore, 2-PL model with fixed thresholds was the best fit model for cultural awareness of teachers.

Items 47 and 48 had the lowest 1-2 threshold parameters, which show that teachers are less likely to strongly disagree with these items than item 46. Items 47 and 48 had the lowest 3-4 threshold parameters showing that teachers are more likely to strongly agree with these items when compared to item 46. The discrimination parameters for the 3 items ranged from (1.070 to 0.487). Item 47 (I believe there are times when “racial statements” should be ignored) had the highest discrimination parameter which means that this item contributed the most towards the latent trait, cultural awareness of teachers. Item 48 had a very low discrimination parameter and this item almost dichotomizes the responses as can be seen from the item characteristic curves in Figure 27. Item 46 had a low discrimination parameter as well, which means that teachers are more likely to strongly agree or strongly disagree with this item rather than have a neutral view with this item. However, this factor pattern coefficient had the highest factor pattern coefficient (0.719).

Table 22

Item Thresholds and Discrimination Parameters for Models 1 through 5-5 (Factor VII)

Items	Parameters	1-PL (model 1)	2-PL (model 2)	1-PL (model 3)	2-PL (model 4)	gender cov. (model 5)	Ethnicity cov. (model 6)	Exp. cov. (model 7)
Item 46	Threshold 1-2	-2.780	-2.880	-2.669	-3.230	-2.879	-3.264	-2.851
	Threshold 2-3	-1.780	-1.847	-1.840	-2.250	-1.843	-2.230	-1.815
	Threshold 3-4	-0.304	-0.335	-0.294	-0.374	-0.331	-0.720	-0.303
	Discrimination	0.751	0.864	0.753	1.182	0.865	0.848	0.865
Item 47	Threshold 1-2	-3.377	-3.446	-1.993	-2.000	-3.446	-3.740	-3.425
	Threshold 2-3	-2.377	-2.413	-1.061	-1.080	-2.410	-2.706	-2.389
	Threshold 3-4	-0.901	-0.901	0.185	0.184	-0.898	-1.196	-0.877
	Discrimination	0.751	1.070	0.753	0.875	1.071	1.049	1.070
Item 48	Threshold 1-2	-3.460	-3.646	-2.412	-2.200	-3.643	-4.212	-3.603
	Threshold 2-3	-2.460	-2.613	-1.187	-1.080	-2.607	-3.178	-2.567
	Threshold 3-4	-0.984	-1.101	0.474	0.435	-1.095	-1.668	-1.055
	Discrimination	0.751	0.487	0.753	0.517	0.487	0.456	0.488
Model	Log Lkhd*	-3118.2	-3094.8	-3097.9	-3089.2	3094.8	-3078.4	-3094.8
	BIC¥		40.82	-2.83	-17.35§§	-6.02 ***(N)***	26.90 (***Y***)	-5.95 (***N***)
	Random cov. Effect	--	--	--	--	0.003	-0.085	0.020
	P**	--	*0.00*	*0.00*	*0.00*	*0.97* ***(N)***	*0.00* ***(Y)***	*0.785* ***(N)***

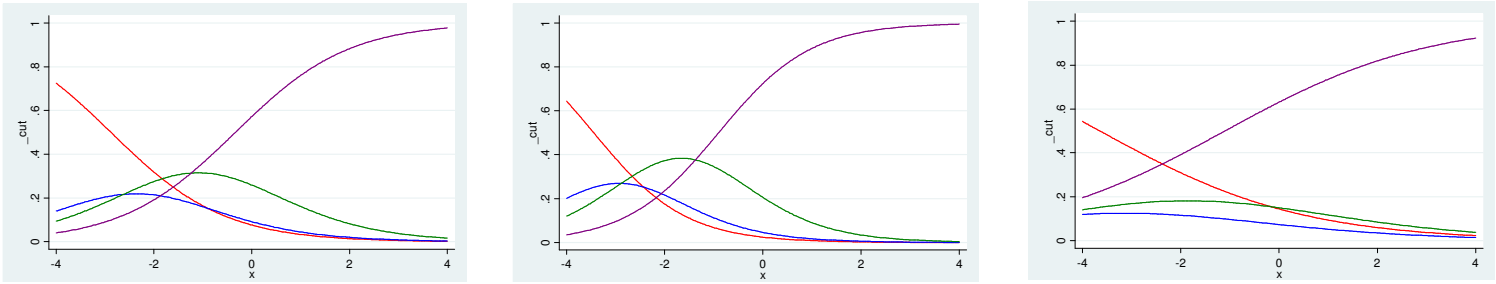
Note: * Log Lkhd represents the Log Likelihood of the model; BIC¥ represents the BIC of current model over previous model; ** P represents the probability that the previous model fits better than the current model; §§ represents the probability of model fit of model 4 over model 2; ***Models 5, 6, and 7 were estimated by including the respective covariates on model 2.***

The covariates, gender, ethnicity, and teaching experience were included in the 2-PL model with fixed thresholds to estimate the effect of these covariates on the latent trait, cultural awareness. Table 22 further shows that when the ethnicity of the teachers was included in the model there was a statistically significant improvement in the model fit as shown by the BIC and p-values. Gender and teaching experience of the teacher did not have a statistically significant impact on the cultural awareness of the teachers. The estimates of overall and differential effects of teachers' ethnicity on their cultural awareness and the individual items that constitute cultural awareness are shown in Table 23. When the effect of ethnicity on the individual items were estimated, there was a differential effect of teaching experience on item 48 only (I believe a child should be referred "for testing" if learning difficulties appear to be due to cultural differences). The item characteristic curves of the items by ethnicity for all 3 items in factor 7 are shown in Figures 28-30.

Table 23
Item Thresholds and Discrimination Parameters for Models with Statistically Significant Covariates (Factor VII)

Covariate		Item 46	Item 47	Item 48
Ethnicity	Threshold 1-2	-3.315	-3.621	-4.451
	Threshold 2-3	-2.279	-2.588	-3.415
	Threshold 3-4	-0.769	-1.077	-1.889
	Discrimination	0.833	1.037	0.495
Model	Log Lkhd*	-3077.990	-3077.970	-3069.690
	BIC¥	-5.275	-5.215	0.614
	Random cov. Effect	-0.078	-0.076	-0.095
	Cov. Effect on Item	-0.018	-0.022	0.043
	P**	0.385	0.367	0.010
	Presence of **DIF**	**NO**	**NO**	***YES***

Note: * Log Lkhd represents the Log Likelihood of the model; BIC¥ represents the BIC of current model over previous model; ** P represents the probability that the previous model fits better than the current model



Item 46 **Item 47** **Item 48**

***Figure* 27. Item Characteristic Curves for Factor VII – Cultural Awareness**

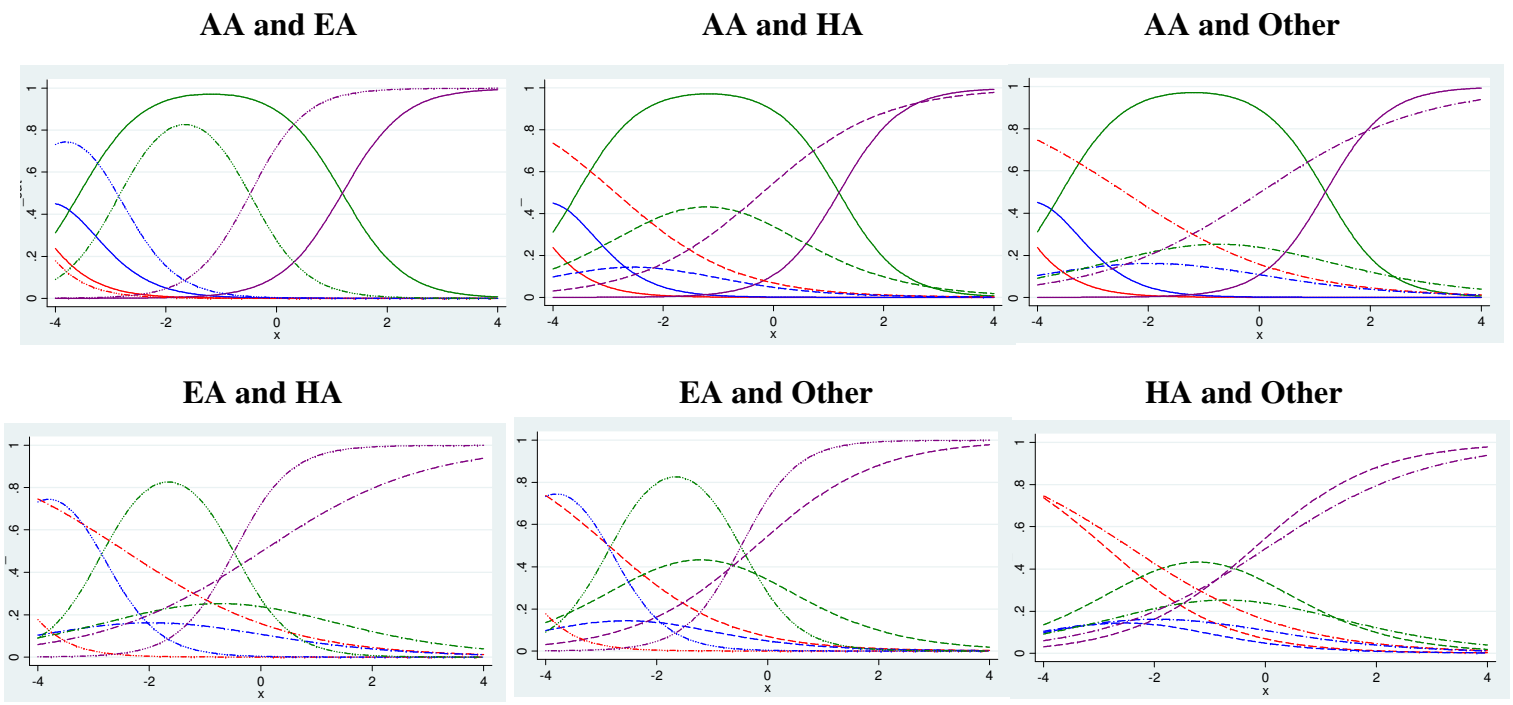


Figure **28. Item Characteristic Curves for Factor VII by Ethnicity – Item 46**

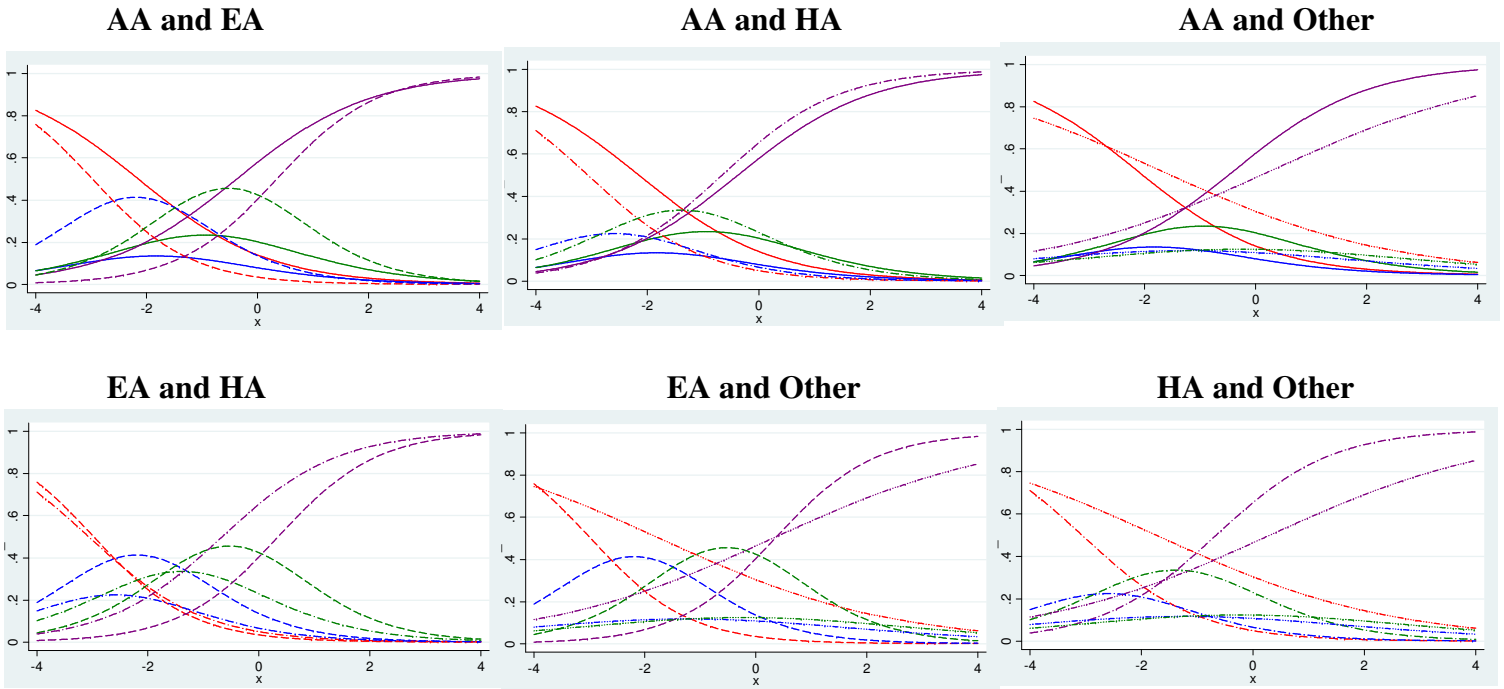


Figure **29. Item Characteristic Curves for Factor VII by Ethnicity – Item 47**

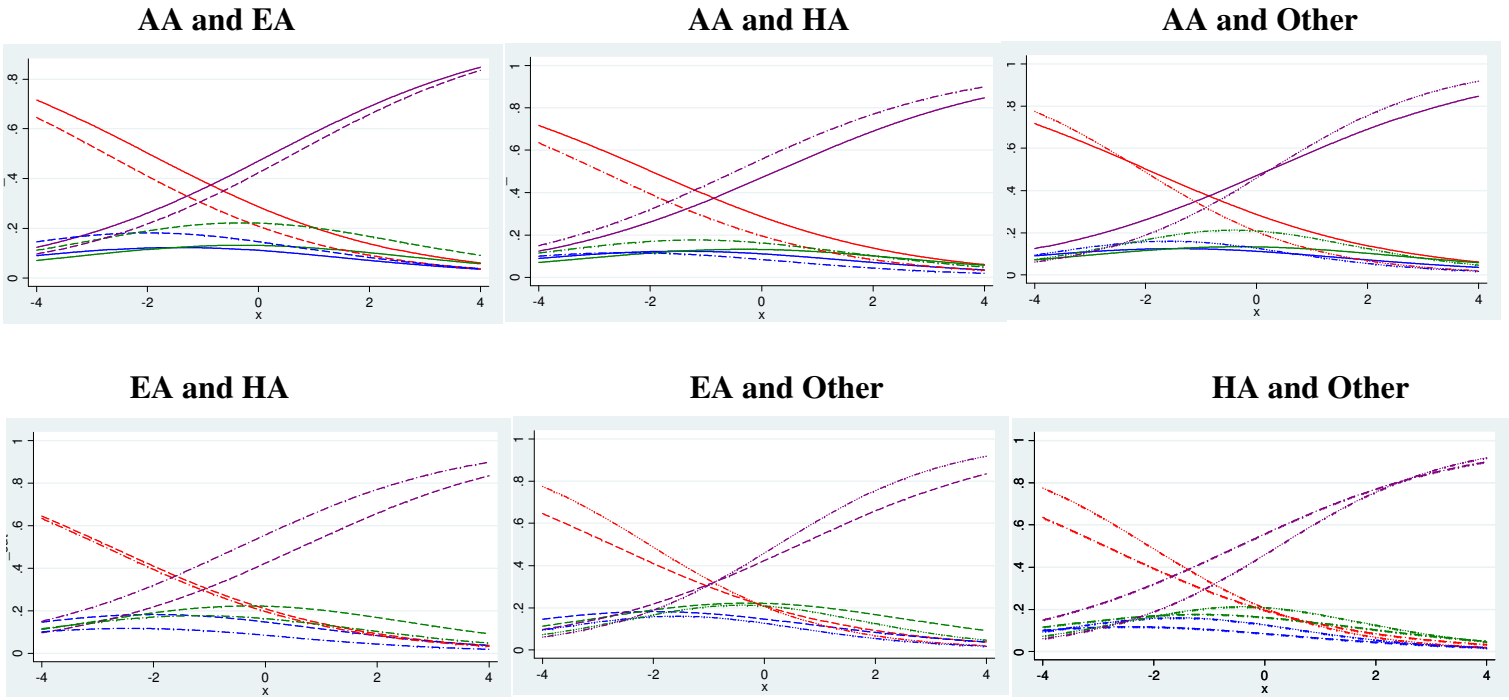


Figure **30. Item Characteristic Curves for Factor VII by Ethnicity – Item 48**

Factor VIII: Teacher Efficacy

The eighth factor, teacher efficacy consisted of 4 items as shown in Table 24. As in the previous factors, Multilevel Item Response theory was applied to this factor, with the latent trait of interest being teacher efficacy. Similar to the previous factors, the analysis was first conducted to investigate the qualities of the items and how much they contribute to teacher efficacy that helps them teach African American students effectively. Then the effects of gender, ethnicity, and teaching experience of the teachers on the latent trait were estimated.

Table 24
Items in Teacher Efficacy Factor

Question No.	Item
23	I believe that some students do not want to learn.
25	I believe there are factors beyond the control of teachers that cause student failure.
49	I believe the teaching of ethnic customs and traditions is not the responsibility of public school personnel.
53	I believe in-service training focuses too much on “multicultural” issues.

The item location (threshold) and slope (discrimination) parameters for the items measuring teacher efficacy for various models are shown in Table 25. After considering the chi-square and the BIC values between the models, the 1-PL IRT model with varying

thresholds was found to be the model of best fit for factor 8. The difference between the BICs of the 1-PL IRT model with varying thresholds and the 2-PL IRT model with fixed thresholds was found to be 175.54. The 2-PL model with varying thresholds did not have a better fit than the 1-PL model with varying thresholds. Therefore, the covariates were included in model 3 to find the direct and the differential effect of gender, ethnicity, and teaching experience on teacher efficacy.

Items 49 and 53 had the lowest 1-2 threshold parameters, which show that teachers are less likely to strongly disagree with these items than the other items in the factor. These items also had the lowest 3-4 threshold parameters showing that teachers are less likely to strongly agree with these items as well when compared to the other items. Items 25 had the highest 1-2, 2-3, and 3-4 threshold parameters indicating that it takes a lot of efficacy of the teachers to strongly agree to this item. Being a 1-PL model, all the items had a discrimination parameter of 0.501. This being a low value for a discrimination parameter suppressed the middle categories of the items. Therefore, almost all the items had only extreme categories (strongly disagree and strongly agree) that were distinguishable. This can be seen from the item characteristic curves in Figure 31. The factor pattern coefficients for these items had similar values ranging from 0.489 to 0.523 except item 53 which had a lower factor pattern coefficient of 0.417. Table 25 further shows that none of the covariates, gender, ethnicity, and teaching experience have a statistically significant impact on teacher efficacy. Therefore, it can be said that teacher efficacy is independent of teaching experience, gender, or ethnicity of the teacher.

Table 25

Item Thresholds and Discrimination Parameters for Models 1 through 5-5 (Factor VIII)

Items	Parameters	1-PL (model 1)	2-PL (model 2)	1-PL (model 3)	2-PL (model 4)	gender cov. (model 5)	Ethnicity cov. (model 6)	Exp. cov. (model 7)
Item 23	Threshold 1-2	-1.097	-1.130	-1.106	-1.112	-1.117	-1.040	-1.059
	Threshold 2-3	0.107	0.134	0.191	0.211	0.090	0.163	0.147
	Threshold 3-4	1.609	1.638	1.320	1.448	1.220	1.290	1.279
	Discrimination	0.499	0.842	0.501	0.714	0.500	0.500	0.501
Item 25	Threshold 1-2	-2.333	-2.400	0.104	0.108	0.003	0.080	0.061
	Threshold 2-3	-1.129	-1.136	1.631	1.656	1.530	1.603	1.587
	Threshold 3-4	0.373	0.368	2.122	2.157	2.020	2.092	2.081
	Discrimination	0.499	0.576	0.501	0.533	0.500	0.500	0.501
Item 49	Threshold 1-2	-0.279	-0.297	-1.838	-1.770	-1.939	-1.863	-1.882
	Threshold 2-3	0.925	0.967	-0.774	-0.746	-0.875	-0.802	-0.818
	Threshold 3-4	2.427	2.471	0.824	0.798	0.724	0.794	0.783
	Discrimination	0.499	0.389	0.501	0.403	0.500	0.500	0.501
Item 53	Threshold 1-2	-0.138	-0.158	-2.055	-1.962	-2.155	-2.080	-2.098
	Threshold 2-3	1.066	1.106	-1.132	-1.078	-1.232	-1.160	-1.175
	Threshold 3-4	2.568	2.610	0.852	0.819	0.752	0.823	0.811
	Discrimination	0.499	0.190	0.501	0.280	0.500	0.500	0.501
Model	Log Lkhd*	-4548.7	-4525.3	-4432.9	-4431.5	-4431.9	-4432.7	-4432.8
	BIC¥		37.83	175.54	-6.08	-7.00 ***(N)***	-8.62(***N***)	-8.73(***N***)
	Random cov. Effect	--	--	--	--	-0.08	-0.006	-0.027
	P**	--	*0.00*	*0.00*	*0.39*	*0.15* ***(N)***	*0.51* ***(N)***	*0.57* ***(N)***

Note: * Log Lkhd represents the Log Likelihood of the model; BIC¥ represents the BIC of current model over previous model; ** P represents the probability that the previous model fits better than the current model

(gender, ethnicity, and experience were fitted on model 3)

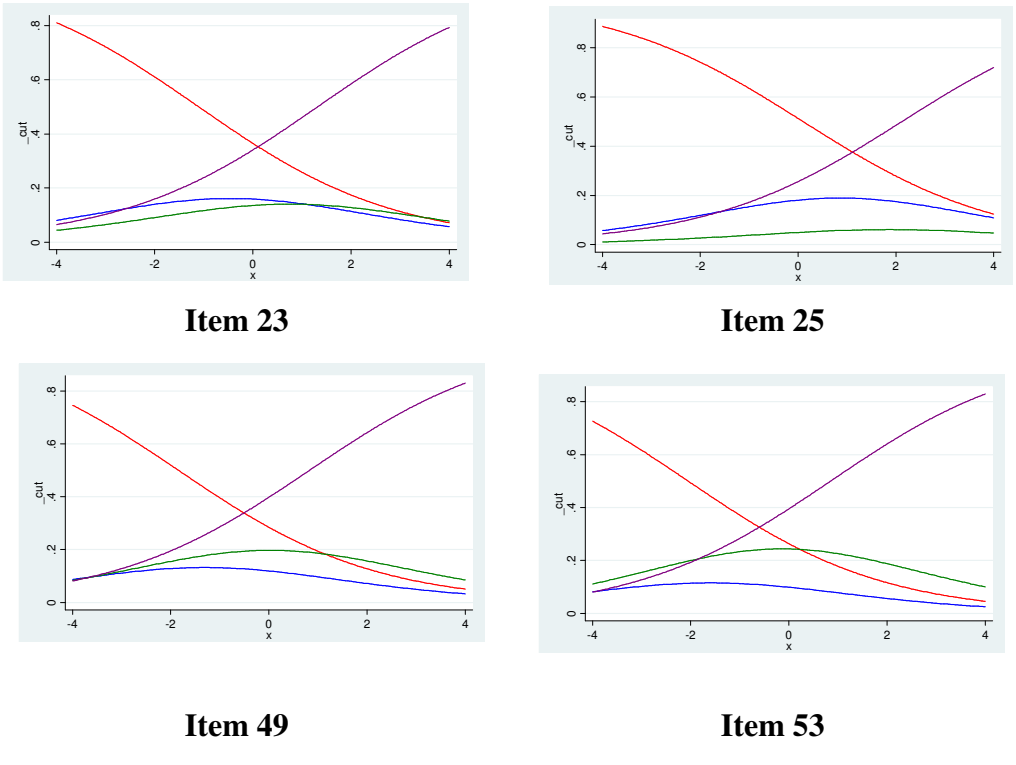


***Figure* 31. Item Characteristic Curves for Factor VIII – Teacher Efficacy**

Discussion

The item analysis shows that items 35, 32, and 31 discriminated between teachers who have stronger beliefs in teaching African American students and teachers with weaker beliefs about teaching African American students better than did the rest of the items. An important implication of this finding is that, if there is a need to develop a shorter instrument, then items with higher discrimination parameters will provide more information and therefore they can be chosen over items with lower discrimination parameters. Teachers tend to pick between strongly agree and strongly disagree when it comes to items 30, 34, 42, 38, and 52. Items 32, 35, 38, and 52 had the least 2-3 threshold

parameters showing that it is easy for teachers with weaker beliefs to crossover between disagree to agree on these items. However, only teachers with extremely positive teacher beliefs endorse items 31, 32, and 34 strongly.

Gender and teaching experience did not significantly affect teacher beliefs on the whole but ethnicity had a statistically significant overall impact on teacher beliefs. The trace lines for the individual items by ethnicity show that African American teachers' perceptions on item 31 differed from the teachers belonging to the rest of the ethnic groups. On item 32, African Americans and European Americans had the maximum difference between them. However, after controlling for the overall effect, item 42, which deals with the teachers having difficulty getting parents from African American families involved in the education of their children was found to have a differential effect between teachers from various ethnic groups. This is a reasonable finding considering that overall, teachers from different ethnic groups communicate differently with parents. Teachers with background similar to that of African American students, such as African American and Hispanic American teachers, are probably able to understand the family dynamics of these students making it easier for them to communicate with the students' families.

The present study demonstrates the use of item response theory, MLIRT in particular, in understanding the latent traits of teachers, teacher beliefs, in this case. The present study also shows the relationship between items and paves way into building item banks that are effective in measuring the cultural beliefs of teachers. Ethnicity plays an important role in shaping up teacher beliefs and this finding can be effectively used in designing teacher induction programs. This analysis can be repeated with other factors to

understand the functioning of those factors, their items, and the effect of various covariates on these factors and items. Additional information such as the characteristics (attitudes towards various factors) of the individual can be gotten by computing the person fit of the analysis, but was beyond the scope of the present study. Furthermore, a comparison between the findings of factor analysis and IRT analysis can be done which seems to be the next logical step in this series of studies.

Comparison between the results yielded by factor analysis and MLIRT analysis is inevitable at this stage. While the results of factor analysis and MLIRT analysis are similar, they are not identical. The factor/pattern coefficients of the items are positively related to the discrimination parameter of the items. However, in factor analysis, the characteristics of the items are determined based on the responses given by the participants while the factor scores for each participant on the factor is calculated based on the item characteristic (factor/pattern coefficient, in this case) and the response given by the person. This leads us back to the circular dependency of person ability (or attitude) and the item characteristic. In the MLIRT analysis, the item characteristics are separated from the person abilities (or attitudes) and the item parameters will remain the same when administered to a similar teacher sample. Moreover, the probability of a person answering in each category (from Strongly Disagree to Strongly Agree) is obtained for each item as shown by the item characteristic curves which indicate the function of the number of categories of the items.

When the discrimination parameter is high, almost all categories are dominant and are distinguishable from each other by the respondents (e.g., items 32 and 35). Therefore,

the number of categories in these items seems to provide additional information about the item. On the contrary, the respondents do not distinguish well between categories for items with low discrimination parameters (e.g., 30 and 34). However, it would be difficult to decide without any further analysis if addition of categories will provide additional information or deletion of categories will result in loss of information.

In sum, although factor analysis provides valuable information, the use of MLIRT on such datasets provides additional information that can help identify the item characteristics free from the attitude level of the respondent, the probability of answering in each category for each item, and also identify the gender/ethnicity/teaching experience biases of items. These are some key findings that can help modify the instrument to a shortened version and yet maximize the information, to be free from biases, and to identify the number of categories that would yield maximum information about the sample. Therefore, both factor analysis and MLIRT provide valuable but different information which answer different questions. However, MLIRT involves complex computation and large sample sizes which may persuade a researcher to prefer factor analysis over MLIRT.

Although the advantages of IRT have been presented, it is necessary to warn the naïve beginner in IRT about the debates surrounding the concept. There have been several critics and criticisms of IRT (e.g., Burton, 2004, 2005; Fan, 1998; Lawson, 1991; MacDonald & Paunonen, 2002). When items do not fit a given statistical model, they are usually rejected in IRT and this leads to that particular area of knowledge not being tested (Burton, 2004). Items with test scores of all zero or all correct are excluded from the

analysis in IRT. However, this is the case in classical analysis as well, because in classical analysis, items with all zeros or all ones will have no variance and therefore, not contribute anything to the analysis. In IRT, such items are excluded in the first phase of the analysis. A valid observation about IRT is the need for large sample sizes, which is not always the case, especially in the behavioral sciences. IRT involves complex computations and is sometimes time-intensive (Burton, 2004, 2005; Lawson, 1991; MacDonald & Paunonen, 2002).

However, the advantages IRT provides over CTT makes the effort worthwhile. Unlike CTT, IRT does not require item error variance to be equal across populations (Raju, Laffitte, & Byrne, 2002). In most practical situations it is almost impossible to find items with equal error variances which make the results of classical test analyses much less robust (Byrne, 1994a, 1994b, 1998, 2001). In IRT, the error variance differs as a function of the person ability (or attitude level) and therefore, can vary from person to person (Hambleton, Swaminathan, & Rogers, 1991). This is an important advantage, especially when using IRT to detect DIF.

While I have demonstrated the use of IRT in measuring the cultural beliefs of urban teachers, I also acknowledge and appreciate the wealth of information classical analyses provide. Instead of taking sides and choosing between the two worlds, I advocate the choice of methodology that is appropriate for the question that drives the research. If used with caution, these two worlds can complement each other and help understand the human behavior and its latent traits, which is the crux of behavioral research.

CHAPTER V

SUMMARY AND DISCUSSION

Summary

The simulation study was conducted to test for the performance of 2-PL MLIRT models under various conditions (test length, sample size, correlation between the predictor variable and the person ability (or attitude), and the distribution of the binomial predictor variable). The substantiation study was conducted to illustrate the contribution of 2-PL MLIRT models to such behavioral studies that involve measurement of attitudes, abilities, or latent traits. The study using the urban dataset illustrated the measurement of item characteristics that form each scale and further demonstrated the advantages of using 2-PL MLIRT models over classical analysis and 1-PL models. The simulation study paves way to help interpret the results of the substantiation study with more confidence as discussed in the forthcoming section.

The results from the simulation study show that test length and sample size have the most effect on the accuracy of the threshold and parameter estimates. A test length of 15 items performs reasonably well with a sample size as low as 200 when sufficient number of iterations (>500) were run and when a predictor variable was present. Therefore, it could be safely said that for the urban data set with a sample size of over 1000 teachers, the estimates obtained were sufficiently accurate, especially because the factors contained lesser number of items. This can be equated to having shorter tests with larger sample sizes which is a win-win situation.

The presence of predictor variables such as ethnicity, gender, and teaching experience further added to the information about the latent traits of interest. The proportion of participants who were male to those who were female is about 0.25. The simulation study helps validate the results of the urban district study because it can be said with confidence that there were no statistically significant difference between the cultural awareness and beliefs of male teachers when compared to that of female teachers. This lack of difference did not happen because of the lack of comparable sample sizes because the results of the first study confirm that unequal sample sizes (or binomial distribution) have negligible effect on the estimates of the parameters.

The urban district study found that the ethnicity had a statistically significant impact on the teacher beliefs factor about African American students, culturally responsive management, and cultural awareness. Teaching experience had a statistically significant effect on culturally responsive management, home and community support, and curriculum and instructional strategies. The gender of the teacher did not have a significant impact on any of the factors that explained teachers' perceptions about cultural awareness and beliefs in teaching African American students.

Discussion

The main ideas that were explained through the course of this dissertation are revisited below:

(1) IRT models can be successfully used to measure the characteristics of items

As the urban district study demonstrated, IRT can be successfully used to measure the characteristics of items that measure the attitudes of respondents. The threshold

parameters of the items indicate the endorsement rate of the items by the respondents. Higher threshold values of an item indicate that teachers who choose more positive categories (such as agree and strongly agree) possess higher attitude levels on the corresponding factor (that the item belongs to) when compared to teachers that choose more positive categories on items with lower threshold values. Higher discrimination parameters indicate the contribution of the item to the overall scale which is discussed in detail in the proceeding sections.

(2) Two-Parameter MLIRT models offer advantages over classical analyses

Unlike classical analyses, 2-PL MLIRT models do not require equal or even comparable sample sizes to test for difference between groups as can be seen from the simulation study. The ratio between group sizes can vary anywhere from being 50:50 to 10:90. This is certainly advantageous, especially because the independent variables of the respondents (e.g., demographics) are usually not controlled by the researcher. For example, the urban district study had 20% male and 80% female teachers, which is generally the case in K-12 teaching population. In cases such as these, it is very necessary for the group sizes to be equal in classical analyses to perform any comparison between groups. Furthermore, the correlation between the predictor variable and the attitude of the respondent does not affect the accuracy or precision of the estimates of item characteristics. Therefore, it is not necessary to perform any additional analysis to remove the effect of the predictor variable before conducting a 2-PL MLIRT analysis. Although the results from factor analysis and MLIRT analysis are comparable, they are not

identical. MLIRT analysis provides more information about the characteristics of items and can therefore be used to build more efficient instruments.

(3) Two-Parameter MLIRT models offer several statistical advantages over 1-PL models and ordinary IRT models

Two-parameter MLIRT models measure the item discrimination parameter which contributes to the amount of information that is explained by a particular item. This can be considered analogous to the "variance explained" statistic in classical analysis. The urban district study illustrates the importance of using 2-PL models because, as it can be seen, when the discrimination parameter is set to be the same for all items, the information about the performance of the categories is lost. When every item has the same slope, it is almost impossible to determine whether the number of categories are adequate or if additional categories have to be included so that the respondents can choose the category that measures their attitude more accurately. The restriction for test length and sample size in 2-PL MLIRT model is much less when compared to the recommended sample size and test length in ordinary IRT models. This gives advantage to the researcher, especially while using real datasets because of the costs and efforts involved in data collection.

Scope of Further Research

The field of MLIRT models provides several opportunities for furthering the research in this area. The simulation study can be extended for cross-classified models, for predictor variables that have more than two categories, and continuous predictor variables. Studies could also be conducted to include several predictor variables and

investigate the interaction effects between the predictors. The advantage of being a part of multilevel modeling is the scope for inclusion of more complicated models. Furthermore, the effect of predictor variables at different levels (item and person level) can be investigated as well.

The substantiation study would hopefully open the field of MLIRT to a wider range of applications for the concept. The effect of level taught by the teachers on their perceptions of cultural awareness and beliefs can be investigated. The instrument can be further developed so that it can be used for adaptive testing to cater to wider audience at both the national and international levels. Urban teachers could be compared to teachers from sub-urban and rural districts to further gain insight into the home environment and the school environment of teachers. The number of categories in the instrument can be modified to find the effect of adding or deleting categories to such instruments.

REFERENCES

Adams, D. L. (Ed.). (1995). *Health issues for women of color: A cultural diversity perspective*. Thousand Oaks, CA: Sage Publications.

Adams, R. J., Wilson, M. & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, *22*, 47-76.

Allen, B. A., & Boykin, A. W. (1992). African American children and the educational process: Alleviating cultural discontinuity through prescriptive pedagogy. *School Psychology Review, 27*, 586-596.

Andersen, E. B. (1973). Conditional inference for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology*, *26*, 31-44.

Anderson, J. A. (2004). The historical context for understanding the test score gap. *Journal of Public Management and Social Policy*, *10*, 1-35.

Andrich, D. (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, *2*, 581-594.

Aronson, E., & Gonzalez, A. (1988). Desegregation, jigsaw, and the Mexican-American experience. In P. A. Katz & D. A. Taylor (Eds.), *Eliminating racism: Profiles in controversy*. New York: Plenum Press.

Ashton, P., & Webb, R. (1982). *Teachers' sense of efficacy: Toward an ecological model.* Paper presented at the annual meeting of the American Educational Research Association, New York.

Bailey, C. T., & Boykin, A. W. (2001). The role of task variability and home contextual factors in the academic performance and task motivation of African American elementary school children. *Journal of Negro Education, 70*, 84-95.

Baker, F. B. (1985). *The basics of item response theory*. Portsmouth, NH: Heinemann.

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques.* New York: Marcel Dekker.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.

Bandura, A. (1995). Exercise of personal and collective efficacy in changing societies. In A. Bandura (Ed.), *Self-efficacy in changing societies* (pp. 1-45). Cambridge, MA: Cambridge University Press.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: W. H. Freeman and Company.

Banks, J. A. (1988). Ethnicity, class, cognitive, and motivational styles: Research and teaching implications. *Journal of Negro Education, 54*, 452-467.

Banks, J. A. (1991a). A curriculum for empowerment, action, and change. In C. E. Sleeter (Ed.), *Empowerment through multicultural education* (pp. 125-141). Albany, NY: State University of New York Press.

Banks, J. A. (1991b). *Teaching strategies for ethnic studies* (5th ed.). Boston: Allyn & Bacon.

Banks, J. A. (1993a). Multicultural education: Historical development, dimensions, and practice. In L. D. Hammond (Ed.), *Review of research in education* (pp. 3-49). Washington, DC: American Educational Research Association.

Banks, J. A. (Ed.). (1993b). *Integrating curriculum with ethnic content: Approaches and guidelines*. Boston: Allyn & Bacon.

Banks, J. A. (1994a). *An introduction to multicultural education*. Boston: Allyn & Bacon.

Banks, J. A. (1994b). Transforming the mainstream curriculum. *Educational Leadership*, 4-8.

Banks, J., A. (1997). *Educating citizens in a multicultural society*. New York: Teacher's College Press.

Banks, C. A. M., & Banks, J. A. (1995). Equity pedagogy: An essential component of multicultural education. *Theory into Practice, 34*, 152-158.

Banks, J. A., & Banks, C. A. (Eds.). (2001). *Multicultural education: Issues and perspectives* (4th ed.). Needham Heights, MA: Allyn & Bacon.

Beretvas, S. N., & Kamata, A. (2005). The multilevel measurement models: Introduction to the special issue. *Journal of Applied Measurement*, *6*, 247-254.

Beretvas, S. N., Meyers, J. L., & Rodriguez, R., A. (2005). The cross-classified multilevel measurement model: An explanation and demonstration. *Journal of Applied Measurement*, *6*, 322-341.

Beretvas, S. N. & Williams, N. J. (2004). The use of hierarchical generalized linear model for item dimensionality assessment. *Journal of Educational Measurement*, *41*, 379-395.

Bielinski, J., & Davison, M. L. (2001).A sex difference by item difficulty interaction in multiple choice mathematics items administered to national probability samples. *Journal of Educational Measurement*, *38*, 51-77.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick, *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29-51.

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement*, *25*, 275-85.

Bock, R. D., & Mislevy, R. J. (1989). *Duplex design: Giving students a stake in educational assessment*. Chicago: Methodology Research Center, NORC.

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden & R. Hambleton (Eds.), *Handbook of modern Item Response Theory* (pp. 438-448). New York: Springer.

Boykin, A. W. (1983). The academic performance of Afro-American children. In J. Spence (Ed.), *Achievement and achievement motives* (pp. 324-337). San Francisco: Freeman.

Boykin, A. W. (1991). The challenges of cultural socialization in the schooling of African American elementary school children: Exposing the hidden curriculum. In J. L. W. Watkins, & V. Chou (Ed.), *Race and education: The roles of history*

and society in educating African American students. Newton, MA: Allyn and Bacon.

Boykin, A. W. (2001). The challenges of cultural socialization in the schooling of African American elementary school children: Exposing the hidden curriculum. In W. Watkins, J. Lewis, & V. Chou (Eds.), *Race and education: The roles of history and society in educating African American students* (pp. 190–199). Boston: Allyn & Bacon.

Boykin, A.W., Albury, A., Tyler, K. M., Hurley, E. A., Bailey, C. T., & Miller, O. A. (2005). Culture-based perceptions of academic achievement among low-income elementary students. *Cultural Diversity and Ethnic Minority Psychology*, *11,* 339-350.

Boykin, A. W., & Allen, B. A. (1988). Rhythmic movement facilitation of learning in working-class Afro-American children. *The Journal of Genetic Psychology, 149*, 335-347.

Boykin, A. W., Allen, B. A., & Davis, L. H. (1997). Task performance of Black and White children across levels of presentation variability. *The Journal of Psychology, 131*, 427-437.

Boykin, A. W., & Cunningham, R. (2001). The effects of movement expressiveness in story content and learning context on the analogical reasoning performance of African American children. *Journal of Negro Education, 70*(1-2), 72-83.

Boykin, A. W., & Toms, F. (1985). Black child socialization: A conceptual framework. In H. P. McAdoo & J. L. McAdoo (Eds.), *Black children: Social, educational, and parental environments*. Beverly Hills, CA: Sage.

Bradford, L., Pitts, V., & Collins, A. (2002). A descriptive study of preservice teachers' perceptions of African American students' ability to achieve in mathematics and science. *The Negro Educational Review*, *53*, 31-42.

Brown, G. T. L. (2004). Teachers' conceptions of assessment: Implications for policy and professional development. *Assessment in Education: Policy, Principles and Practice*, *11*(3), 305-322.

Bruno, J. E. & Dirkzwager, A. (1995). Determining the optimal number of alternatives to a multiple-choice test item: An information theoretic perspective. *Educational and Psychological Measurement*, *55*, 959-966.

Burton, R. (2004). Can item response theory help us improve our tests? *Medical Education*, *38*, 338-339.

Burton, R. (2005). Multiple-choice and true/false tests: Myths and misapprehensions. *Assessment and Evaluation in Higher Education*, *30*(1), 65-72.

Byrne, B. M. (1994a). *Structural equation modeling with EQS and EQS/Windows: Basic concepts, applications, and programming*. Thousand Oaks, CA: Sage.

Byrne, B. M. (1994b). Testing for the factorial validity, replication, and invariance of a measuring instrument: A paradigmatic application based on the Maslach Burnout Inventory. *Multivariate Behavioral Research*, *29*, 289-311.

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Erlbaum.

Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Mahwah, NJ: Erlbaum.

Cantrell, C.E. (1999). Item response theory: Understanding the one-parameter Rasch model. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 5, pp. 171-192). Stamford, CT: JAI Press.

Carter, N. P., Hawkins, T. N., & Natesan, P. (Forthcoming). The impact of verve on the academic achievement of African American and European American middle school students. *The Journal of Educational Foundations.*

Cohen, E. G. (1994). *Designing groupwork: Strategies for the heterogeneous classrooms* (2nd ed.): New York: Teachers College Press.

Cohen, E. G., & Roper, S. S. (1972). Modification of interracial interaction disability: An application of status characteristics theory. *American Sociological Review, 37*, 643-657.

Cox, E. P. (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research*, *17*, 407-422.

Davey, T., Nering, M. L., & Thompson, T. (1997). *Realistic simulation of item response data*. ACT research report series 97-4, Iowa City, IA: ACT.

Delpit, L. (1995). Other people's children: culture conflict in the classroom. New York: The New Press.

Digest of Education Statistics. (2003). *Percent of the population 3 to 34 years old enrolled in school, by race/ethnicity, sex, and age: Selected years, 1980 to 2003*. Retrieved from the National Center for Educational Statistics Website: http://nces.ed.gov/programs/digest/d04/tables/dt04_006.asp on April 20, 2006.

Dill, E., & Boykin, A. W. (2000). The comparative influence of individual, peer tutoring, and communal learning contexts on the text recall of African American children. *Journal of Black Psychology, 26*(1), 65-78.

Edward, C. W. (1993). *Revising SAT verbal items to eliminate Differential Item Functioning*. College Board Report, 93-2, New York: The College Board.

Epstein, J. L. (1987). Toward a theory of family-school connections: Teacher practices and parent involvement. In K. Hurrelmann, F. Kaufman and F. Loel (Eds.), *Social Intervention: Potential and Constraints* (pp. 121-136). New York : Walter de Gruyter.

Epstein, J. L. (1996). Perspectives and previews on research and policy for school, family, and community partnerships. In A. Booth & J. Dunn (Eds.), *Family-school links: How do they affect educational outcomes?* (pp. 209-246). Mahwah, NJ: Lawrence Erlbaum.

Epstein, J. L., & Hollifield, J. H. (1996). Title I and school-family-community partnerships: Using research to realize the potential. *Journal of Education for Students Placed at Risk*, *1*, 263-278.

Esposito, C. (1999). Learning in urban blight: School climate and its effect on the school Performance of urban, minority, low income children. *School Psychology Review, 28*, 365-377.

Fan, X. (1998). Item response theory and classical test theory: A comparison of their item/person statistics. *Educational and Psychological Measurement, 58*, 357-381.

Fan, X., Felsövályi, A., Sivo, S. A., & Keenan, S. C. (2002). *SAS for Monte Carlo studies: A guide for quantitative researchers*, Cary, NC: SAS Institute Inc.

Fast Facts Aldine Independent School District. *Aldine Independent School District*. Retrieved February 2, 2007, from http://www.aldine.k12.tx.us/district_info/fast_facts.cfm.

Ford, D.Y., Grantham, T.C., & Harris III, J.J. (1998). Multicultural gifted education: A wake-up call to the profession. *Roeper Review*, *19*, 72-78.

Fordham, S. (1988). Racelessness as a strategy in Black students' school success: Pragmatic strategy or pyrrhic victory? . *Harvard Educational Review, 58*, 54-84.

Fordham, S. (1999). Dissin' "the standard": Ebonics as guerrilla warfare at Capital High. *Anthropology & Education Quarterly, 30*(3), 272-293.

Foster, M. (1992). Sociolinguistics and the African American community: Implications for literacy. *Theory into Practice, 31*, 303-311.

Fox, J.–P. (2004). Applications of multilevel IRT modeling. *School Effectiveness and School Improvement*, *15*, 261-280.

Fox, J.-P. (2005). Multilevel IRT using dichotomous and polytomous response data. *British Journal of Mathematical and Statistical Psychology*, *58*, 145-172.

Fox, J.-P., & Glas, C.A.W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*, 269–286.

Fox, J.-P., & Glas, C.A.W. (2003). Bayesian modeling of measurement error in predictor variables using item response theory. *Psychometrika*, *68*, 169–191.

Garcia, E. (2001). Educating Mexican American students: Past treatment and recent developments in theory, research, policy, and practice. In J. A. Banks & C.A.M. Banks *Handbook of research on multicultural education* (pp. 372-387). San Francisco: Jossey-Bass.

Gay, G. (1995). Mirror images on common issues: Parallels between multicultural education and critical pedagogy. In C. E. Sleeter & P. L. McLaren (Eds.), *Multicultural education, critical pedagogy, and the politics of difference* (pp. 155-190). Albany, NY: State University of New York Press.

Gay, G. (2000). *Culturally responsive teaching: Theory, research and practice*. New York: Teachers College Press.

Gay, G. (2002). Culturally responsive teaching in special education for ethnically diverse students: Setting the stage. *Qualitative Studies in Education*, *15*, 613-629.

Gibson, S., & Dembo, M. H. (1984). Teacher efficacy: A construct validation. *Journal of Educational Psychology, 76*, 569-582.

Green, R. S. (2001, Winter). Closing the achievement gap: Lessons learned and challenges ahead. *Teaching and Change*, *5*, 15-24.

Gregory, A & Mosely, P. (2004). The discipline gap: Teacher's views on the over-representation of African American students in the discipline system. *Equity &*

Excellence in Education, *37*, 18-30.

Hale-Benson, J. E. (1986). *Black children: Their roots, culture, and learning styles.* (2nd ed.). Baltimore, MD: John Hopkins University Press.

Halpin, A. W. & Croft, D. B. (1963). *The organizational climate of schools*. Chicago: Midwest Administration Center of the University of Chicago.

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer Nijhoff.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hammersley, J. M., & Handscomb, D. C. (1964). *Monte Carlo methods*. Methuen, London.

Hays, R.D., Morales, L.S., & Reise, S.P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care, 38,* 28-42.

Henard, D. (2000). Item response theory. In L. Grimm & P. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 67-98). Washington, DC: American Psychological Association.

Henry, G. (1986). *Cultural diversity awareness inventory=Invetorio sobre el reconocimiento de diversas culturas*. Hampton, VA: Hampton University. Mainstreaming Outreach Project. (ERIC Document Reproduction Service No.ED 282657).

Hilliard, A. (1992). Behavioral style, cultural, and teaching and learning. *The Journal of Negro Education*, *61* 370-377.

Hollins, E. R., & Spencer, K. (1996). Restructuring schools for cultural inclusion: Changing the schooling process for African American youngsters. *Journal of Education, 172*(2), 89–100.

Howard, T. (2001). Powerful pedagogy for African American students: A case of four teachers. *Urban Education, 36*, 179 –202.

Hoy, W.K., & Miskel, C. G. (2005). *Educational administration: Theory into practice*. (7th ed) New York: McGraw-Hill.

Hulin, C. L., Lissak, R. I., & Drasgow, R. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement*, *6*, 249-260.

Hurley, E. A., Allen, B. A. & Boykin, A.W. (2005). Communal vs. individual learning of a math-estimation task: African American children and the culture of cooperative learning contexts. *Journal of Psychology: Interdisciplinary & Applied*, *139*, 513-527.

Irvine, J. (1990). *Black students and school failure*. New York: Praeger.

Irvine, J. J. (2003). *Educating teachers for a diverse society: Seeing with the cultural eye*. New York: Teachers College Press.

Irvine, J. J., & Armento, B. (2001). *Culturally responsive teaching: Lesson planning for elementary and middle grades*. New York: McGraw-Hill.

Jencks, C., & Phillips, M. (1998). The Black-White test score gap: An introduction. In C. Jencks & M. Phillips (Eds.), *The Black-White test score gap* (pp. 1-51). Washington, DC: Brookings Institution.

Kamakura, W.A., & Balasubramanian, S.K. (1989). Tailored interviewing: An application of item response theory for personality measurement. *Journal of Personality Assessment, 53*, 502-519.

Kamata, A. (1998). *Some generalizations of the Rasch Model: An application of the Hierarchical Generalized Linear Model*. Unpublished doctoral dissertation. Michigan State University, Ann Arbor.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*, 79-93.

Kelley, R., Thornton, B., & Daugherty, R. (2005). Relationships between measures of leadership and school climate. *Education*, *126*, 17-25.

King, J. E. (1992). Diaspora literacy and struggle against miseducation in the Black community. *The Journal of Negro Education, 61*, 317-340.

King, J. E. (2004). Culture-centered knowledge: Black studies, curriculum transformation, and social action. In J.A. Banks & C.A. Banks, *Handbook of research: Multicultural education* (pp. 628-634). San Francisco: Jossey-Bass.

Kulick, E., & Hu, P.G. (1989). *Examining the relationship between differential item functioning and item difficulty*. College Board Report, 89-5, New York: The College Board.

Ladson-Billings, G. (1990). Like lightning in a bottle: Attempting to capture the pedagogical excellence of successful teachers of Black students. *Qualitative Studies in Education, 3*, 335-344.

Ladson-Billings, G. (1994). *The dreamkeepers: Successful teachers of African American children.* San Francisco: Jossey-Bass.

Ladson-Billings, G. (Ed.). (1995). *Multicultural teacher education: Research, practice, and policy.* New York: Macmillan.

Larke, P.J. (1990) Cultural diversity awareness inventory: Assessing the sensitivity of pre-service teachers. *Action in Teacher Education*, *12*, 23 – 30.

Lawson**,** S. (1991). One parameter latent trait measurement: Do the results justify the effort? In B. Thompson (Ed.), *Advances in educational research: Substantive findings, methodological developments* (Vol. 1, pp. 159-168). Greenwich, CT: JAI.

Le, V. (1999). *Identifying DIF on the NELS 88: History achievement test*. Report for CRESST, CSE- TR-511, Washington, DC: Office of Education Research and Improvement.

Levine, M. V. (1984). *An introduction to multilinear formula score theory*. Model-Based Measurement Laboratory Report 84-4. University of Illinois. Urbana.

Lipman, P. (1995). "Bringing out the best in them": The contribution of culturally relevant teachers to educational reform. *Theory into Practice*, *34*, 202-209.

Lord, F. M. (1968). *Some test theory for tailored testing*. Princeton, NJ: Educational Testing Service.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Love, A. (2001). *Teachers' beliefs and their relationship to student achievement in two African American urban schools*. Digital Dissertation Abstracts. (UMI No. 030012).

Maier, K. S. (2001). A Rasch hierarchical measurement model. *Journal of Educational and Behavioral Statistics*, *26*, 307–330.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*, 187-212.

Mbiti, J. S. (1970). *African religions and philosophy*. New York: Doubleday.

McDonald, P., & Paunonen, S. V. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement*, *62*, 921-943.

McKinley, R., & Mills, C. (1989). Item response theory: Advances in achievement and attitude measurement. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 1, pp. 71-135). Greenwich, CT: JAI Press.

Merriam-Webster Inc. (1994). *Merriam-Webster's dictionary of English usage*. Springfield, MA: Merriam-Webster Inc.

Miller, S. M., Miller, K. L., & Schroth, G. (1997). Teacher perceptions of multicultural training in preservice programs. *Journal of Instructional Psychology*, *24*, 222-232.

Milner, R (2005). Stability and change in US prospective teachers' beliefs and decisions about diversity and learning to teach. *Teaching and Teacher Education, 21*, 767-786.

Milner, R.H., Flowers, L.A., Moore, E.Jr., Moore, J.L., & Flowers, T. A. (2003). Preservice teachers' awareness of multiculturalism and diversity. *High School Journal*, *87*, 63- 80.

Mislevy, R. J. (1983). Item response models for grouped data. *Journal of Educational Statistics*, *8*, 271-288.

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association*, *80*, 993-997.

Mislevy, R. J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement*, *11*, 81-91.

Mislevy, R. J., & Bock, R. D. (1989). A hierarchical item response model for educational testing. In R. Bock (Ed.), *Multilevel analysis of educational data* (pp. 57-74). San Diego, CA: Academic Press.

Moemeka, A. A. (1998). Communalism as a fundamental dimension of culture. *Journal of Communication, 48*(4), 118–141.

Moll, L. & Gonzalez, N. (2004). Engaging life: A funds-of-knowledge approach to multicultural education. In J. A. Banks & C. A. Banks (2nd ed.), *Handbook of research on multicultural education* (pp.699 – 715). San Francisco: Jossey-Bass.

Monroe, C. (2005). Why are "bad boys" always black? Causes of disproportionality in school discipline and recommendation for change. *The Clearing House*, *79*(1), 45-49.

Moon-Merchant, V. & Carter, N. P. (2004). A teacher induction model for urban settings: follow-up study. *Journal of Public Management and Social Policy*, *10*, 39-54.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159-176.

Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden & R. Hambleton (Eds.), *Handbook of modern Item Response Theory* (pp. 153-164). New York: Springer.

Muraki, E., & Bock, R. D. (1998). *PARSCALE (version 3.5): Parameter scaling of rating data*. Chicago, IL: Scientific Software Inc.

National Maternal and Child Health Resource Center on Cultural Competency (NMCHCCC). (1997). *Journey towards cultural competency: Lessons learned.* Vienna, VA: Maternal and Children's Health Bureau Clearinghouse.

Neal, L., McCray, A., Webb-Johnson, G., & Bridgest, S. (2003). The effects of African American movement styles on teachers: Perceptions and reactions. *The Journal of Special Education*, *37,* 49-57.

Neyman, J. & Scott, E.L. (1948). Consistent estimates based on partially consistent observations. *Econometrika*, *16*, 1-5.

Nieto, S. (2000). *Affirming diversity: The sociopolitical context of multicultural education.* New York: Longman.

Nieto, S. (2004). *Affirming diversity: The sociopolitical context of multicultural education (*4th ed.). Boston, MA: Allyn and Bacon.

Ogbu, J. U. (1986). The consequences of the American caste system. In U. Neisser (Ed.), *The school achievement of minority children: New perspectives*. Hillsdale, NJ: Erlbaum.

Ostini, R., & Nering, M. L. (2006). *Polytomous Item Response Theory models*. Thousand Oaks, CA: Sage.

Paley, V. G. (1979). *White teacher*. Cambridge, MA: Harvard University Press.

Pang, V. O., & Sablan, V. A. (1998). Teacher efficacy: How do teachers feel about their abilities to teach African American students? In M. E. Dilworth (Eds.), *In being responsive to cultural differences: How teachers learn* (pp. 39-58). Thousand Oaks, CA: Corwin Press.

Patz, R. J., & Junker, B. W. (1999a). A straight forward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*, 146–178.

Patz, R. J., & Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*, 342–366.

Phuntsog, N. (2001). Culturally responsive teaching: What do selected United States elementary school teachers think? *Intercultural Education*, *12*, 51-64.

Pike, G. R. (1990). *The performance of females and males on the ACT-COMP exam: An analysis of Differential Item Functioning using Samejima's Graded model*. Knoxville, TN: Center for Assessment Research and Development.

Pohan, C. A., & Aguilar (2001). Measuring educator's beliefs about diversity in personal and professional contexts. *American Education Research Journal*, *38*(1), 159-182.

Powers, D. A. & Xie, Y. (2000). *Statistical methods for categorical data analysis*. San Diego, CA: Academic Press.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004a). *GLLAMM Manual*. U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 160, University of California, Berkeley.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004b). Generalized multilevel structural equation modelling. *Psychometrika, 69*, 167-190.

Raftery, A. E. (1986). Choosing models for cross-classifications. *American Sociological Review*, *51*, 145-46.

Raftery, A. E. (1995). Bayesian model selection in social research. In P. Marsden (Ed.), *Sociological methodology* (pp. 111-163). Washington DC: The American Sociological Association.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory, *Journal of Applied Psychology*, *87*, 517-529.

Ramirez, M & Castaneda (1974). *Cultural democracy, bi-cognitive development and education.* New York: Academic Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded Edition, 1980. Chicago: University of Chicago Press).

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Raudenbush, S., Bryk, A., Cheong, Y.F., & Congdon, R. (2004). *HLM 6: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.

Raudenbush, S. W., & Sampson, R. J. (1999). Ecometrics: Toward a science of assessing ecological settings, with application to the systematic social observation of neighbourhoods. *Sociological Methodology*, *29*, 1–41.

Reckase, M. D. (1997). The past and the future of multidimensional item response theory. *Applied Psychological Measurement*, *21*, 25.

Reckase, M. D. & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement*, *15,* 361.

Ree, M. J., & Jensen, H. E. (1980). *Item characteristic curve parameters: Effects of sample size on linear equating*. Brooks Air Force Base, TX: Air Force Human Resources Laboratory, Air Force Systems Command.

Reise, S. P. (2000). Using multilevel logistic regression to evaluate person-fit in IRT models. *Multivariate Behavioral Research*, *35*, 543-568.

Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items?, *Psychological Methods*, *8*, 164-184.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement*, *27*, 133-144.

Rijmen, F., Tuerlinckx, F., & De Boeck, P. (2002). *A nonlinear mixed model framework for IRT models*. (Unpublished manuscript): University of Leuven, Belgium.

Rock, D. A., Pollack, J. M., & Quinn, P. (1995). *Psychometric report for NELS:88 base year through second follow-up*. NCES report 95-382, Washington, DC: U.S. Department of Education.

Rogers, W. T. & Harley, D. (1999). An empirical comparison of three-and four-choice items and tests: Susceptibility to testwiseness and internal consistency reliability. *Educational and Psychological Measurement*, *59*, 234-247.

Rokeach, M. (1968). *Beliefs, attitudes, and values: A theory of organization and change*. San Fransico: Jossey-Bass.

Rosenthal, R. & Jacobson, L. (1968). *Pygmalion in the classroom. Teacher expectation and pupils' intellectual development*. New York: Holt, Rhinehart and Winston.

Rothernberg, J. J., McDermott, P. C., & Gormley, K. A. (1993). A comparison of student teacher and supervisor perceptions of student teaching. *Journal of Education for Teaching*, *19*, 273-278.

Sackney, L. (1988). *Enhancing school learning climate: Theory, research and practice*. (SSTA Research Center Report No.180). Saskatoon, Saskatchewan: Saskatchewan University, Department of Education.

Saklofske, D. H., Michayluk, J. O., & , & Randhawa, B. S. (1988). Teacher efficacy and teaching behaviors. *Psychological Reports, 63*, 407-414.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph*, *No. 17*, *34,* Part 2.

Samejima, F. (1979). *A new family for multiple choice items* (Research Rep. No. 79-4). Knoxville: University of Tennessee.

Samejima, F. (1997). Graded response model. In W.J. van der Linden & R. Hambleton (Eds), *Handbook of modern item response theory* (pp. 85-100). New York: Springer-Verlag.

Sankofa, B. M., Hurley, E. A., Allen, B. A., & , & Boykin, A. W. (2005). Cultural expression and Black students' attitudes toward high achievers. *The Journal of Psychology, 139*(3), 247-259.

Shade, B.J., Kelly, C., & Oberg, M. (1997). *Creating culturally responsive classrooms.* Washington, DC: American Psychological Association.

Singham, M. (1998, September). The canary in the mine: Closing the achievement gap between African American and European American Students. *Phi Delta Kappan*, *6*, 1-33.

Slavin, R. E. (1983). *Cooperative learning*. New York: Longman.

Slavin, R. E. (2001). Cooperative learning and intergroup relations. In J. A. Banks & C. A. M. Banks (Eds.), *Handbook of research on multicultural education* (pp. 628-634). New York: Jossey-Bass.

Sleeter, C. E. (1995). An analysis of the critiques of multicultural education. In J. A. Banks & C. A. M. Banks (Eds.), *Handbook of research on multicultural education* (pp. 81-94). New York: Macmillan.

Sleeter, C. E. (2001). Preparing teachers for culturally diverse schools: Research and the overwhelming presence of whiteness. *Journal of Teacher Education, 52*, 94-106.

Spence, J. T. (1985). Achievement American style: The rewards and costs of individualism. *American Psychologist, 40*, 1285–1295.

Swaminathan, J., & Gifford, J. A. (1983). Estimation of parameters in the three-parameter latent trait model. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 13–30). New York: Academic.

Swartz, E. & Bakari, R. (2005). Development of the teaching in urban schools scale. *Teaching and Teacher Education*, *21,* 829-841.

Taylor, G. (2000). Talking to the dreamkeepers: Culturally responsive teachers reflect on the educational experiences of African American students. *Action in Teacher Research, 21* 101-109.

Texas Education Agency. (2004 - 2005). *Academic Excellence Indicator System*. Retrieved from the Texas Education Agency Website http://www.tea.state.tx.us/perfreport/aeis/ on August 6, 2006.

Thissen, D., Chen, W-H, & Bock, R.D. (2003). *MULTILOG (version 7)* [Computer software]. Lincolnwood, IL: Scientific Software International.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* Washington DC: American Psychological Association.

Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach*. New York: Guilford.

Tucker, L.R., Koopman, R.F., & Linn, R. L. (1969). Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika*, *34*, 421-459.

Tuerlinckx, F., Rijmen, F., Molenberghs, G., Verbeke, G., Briggs, D., Van den Noortgate, W., et al. (2004). Estimation and software. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 343-373). New York: Springer.

Tyler, K. M., Boykin, W. A., & Walton, T. R. (2006). Cultural considerations in teachers' perceptions of student classroom behavior and achievement. *Teaching and Teacher Education, 22*, 998-1005.

Villegas, M., & T. Lucas. (2002.) *Educating culturally responsive teachers.* Albany, New York: State University of New York Press.

Walter-Roberts, P. F., Natesan, P., & Carter, N. P. (in preparation). Development and validation of an instrument to measure and assess the cultural beliefs and awareness of urban teachers.

Warm, T.A. (1978). *A primer of item response theory* (Technical Rep. No. 940279). U.S. Coast Guard Institute, Oklahoma City, OK: U.S. Government Printing Office.

Webb-Johnson, G. & Carter, N. (2005). [*Cultural awareness and beliefs inventory*]. Unpublished data.

Weiss, R. E. (2004). *Modeling longitudinal data*. New York: Springer.

Williams, N.J. (2003). *Item and person parameter estimation using hierarchical generalized linear models and polytomous Item Response Theory models*. Unpublished doctoral dissertation. The University of Texas at Austin.

Wright, B. D., & Linacre, J. M. (1997). *User's guide to BIGSTEPS: Rasch-model computer program*. Chicago, IL: MESA Press.

Wright, B.D., & Stone, M.H. (1979). *Best test designs: Rasch measurement*. Chicago, IL: Mesa Press.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement*, *26*, 55-66.

APPENDIX A

MATLAB Program for Data Generation

```
fprintf(1,'Generating values for all sample sizes and test lengths');
Th=[200 500 1000 2000];
L=[15 30 60]
P=[.1 .25 .4 .5]
R=[.35 .8]
nT=length(Th)
nL=length(L)
nP=length(P)
nR=length(R)
K=4
 for indn=1:nT
  nTh=Th(indn)
  for indL=1:nL
     M=L(indL)
        alpha(1)=.75
        beta(1,1)=unifrnd(-2.8,-1)
        beta(1,2)=unifrnd(-1,1);
        beta(1,3)=unifrnd(1,2.8);
    for i = 2:M
          alpha(i)=alpha(i-1)+((1.33-0.75)/(M-1));
```

```
            beta(i,1)=unifrnd(-2.8,-1);
            beta(i,2)=unifrnd(-1,1);
            beta(i,3)=unifrnd(1,2.8);
     end
       N = 1;
   fprintf(1,'diplaying the simulation parameters \n');
 fprintf('\n number of item(s): %d',M);
fprintf('\n number of categories: %d',K);
fprintf('\n number of ability indeces: %d',nTh);
fprintf('\n number of responses per item per per abiilty index: %d',N);
for m = 1:M
   fprintf(1,'\n \n alpha and betas for item: %d\n',[m])
   fprintf(1,'\t alpha[%d]: %f\n',[m,alpha(m)]);
   for k = 1:K-1
      fprintf(1,'\t \t (b%d)[%d]: %f',[k-1,m,beta(m,k)]);
   end
end
fname = '60.txt';
fprintf(1,'data will be written to %s',fname);
fprintf(1,'\n');
 Y = zeros(N*nTh,M);
X = zeros(N*nTh,M);
```

```
for indP=1:nP
   x=P(indP)
   for indR=1:nR
      y=R(indR)
      [muTh,sigTh,muVec,sigVec,gender,th] = genParams(x,y,nTh);
    %th = linspace(-3,3,nTh);th=th(:);
  for p = 1:M
    fprintf('\t ...generating data for item: %d\n',p);
    [yVec,xVec] = simu(K,nTh,N,alpha(p),beta(p,:),th,1);
    Y(:,p) = yVec;
    X(:,p) = xVec;
  end
  fp = fopen(fname,'wt');
  fprintf(fp,'%d\t',[M,K,nTh,N]);fprintf(fp,'\n');
  fprintf(fp,'%6.4f\t',alpha);fprintf(fp,'\n');
  fprintf(fp,'%6.4f\t',beta);fprintf(fp,'\n');
 for p = 1:size(Y,1);
   fprintf(fp,'%d\t',Y(p,:));
   fprintf(fp,'%6.4f\t',X(p,:));
   fprintf(fp,'\n');
  end
   fclose(fp);
```

```
    bPrior = beta';
    shname = strcat(int2str(M),int2str(nTh),int2str(indP),int2str(indR))
    xlswrite('beta', beta, shname)
% call mat2winbugs
mat2bugs(['dataBugs',int2str(M),int2str(nTh),int2str(indP),int2str(indR),'.txt'],'y',Y(:)+1,'I
',M,'J',K,'N',nTh,'theta',th,'muTh',muVec,'sigTh',sigVec,'ddddfff');
clear Y th muVec sigVec;
%dos('BackBugs14.lnk /par "..\\..\\MATLAB6p5\\toolbox\\winbug\\eg1_src.txt"');
%S=bugs2mat('bugsIndex.txt','bugs1.txt');
        end
      end
    end
  clear beta;
end
```

Genparams Procedure

```
function [mu,sig,muVec,sigVec,gender,th] = genParams(r,p,nTh)
cp = 1-p;
rsq = r^2;
lb = max(0,1-(rsq/p)) % left boundary for sigma2^2
rb = min(1,(1-rsq)/p)
s2 = lb + (rb-lb).*rand(1,1)
s1 = ((1-rsq)-(p*s2))/cp
```

```
mu2 = r*sqrt(cp/p)

mu1 = -mu2*p/cp

mu = cp*mu1 + p*mu2

sig = cp*s1 + p*s2 + cp*(mu1-mu)^2 + p*(mu2-mu)^2

TOL = 1e-6;

if (abs(mu)>TOL)

   error('marginal mean is not zero')

end

if (abs(sig-1)>TOL)

   error('marginal sigma2 is not unity')

end

if ( (s2>1) | (s2<0))

   error('s2 can not be negative or greater than one')

end

if ( (s1>1) | (s1<0))

   error('s1 can not be negative or greater than one')

end

if ( (r>=1) | (r<=-1))

   error('correlation has to be < 1 and > -1')

end

if ( (p>=1) | (p<=0))

   error('proportion has to be between 0 and 1')
```

```
end
%muVec = [mu1,mu2];
%sigVec = [s1,s2];
 % generating gender and corresponding abilities
gender =  binornd(1,p,nTh,1)+1;
muVec = mu2*ones(nTh,1);
sigVec = s2*ones(nTh,1);
muVec(find(gender==1)) = mu1;
sigVec(find(gender==1)) = s1;
 th = normrnd(muVec,sqrt(sigVec));
 mu = [mu1,mu2];
sig = [s1,s2];
```

APPENDIX B

WINBUGS Code for Parameter Estimation

```
model
{
    for( n  in 1:N)
   {
         for( i in 1: I)
         {
            #a[i] < - exp(alpha[i])
            Q[n,i,1]<- 1
            for( j in 1: J-1)
            {
               #logit(Q[n,i,j]) <- a[i,j] + (beta[i]*theta[n,1])

               logit(Q[n,i,j+1]) <- a[i]*(theta[n,1]-b[i,j])
            }
             P[n,i,J]<-Q[n,i,J]
             for(j in 1:J-1)
             {
               P[n,i,j] <- Q[n,i,j]-Q[n,i,j+1]
             }
             y[(i-1)*N+n,1] ~ dcat(P[n,i,1:J]) # likelihood
```

```
        }
    }
    for (n in 1: N)
    {
        prec[n,1] <- 1/sigTh[n,1]
        theta[n,1] ~ dnorm(muTh[n,1],prec[n,1])
    }
    for (i in 1: I)
    {
        a[i] ~ dunif(0.5,1.5)
        b[i,1] ~ dunif(-3,0)
        b[i,2] ~ dunif(b[i,1],1.5)
        b[i,3] ~ dunif(b[i,2],3)
    }
}
```

Sample WINBUGS Procedure to Load the Data and Implement the Model

```
check('C:/Documents and Settings/tlacgrad/Desktop/itm/model.txt')
data('C:/Documents and Settings/tlacgrad/Desktop/itm/dataBugs1520011.txt')
compile(1)
gen.inits()
set(a)
```

```
set(b)

update(1000)

trace(*)

stats(*)

density(*)

history(*)

coda(*,output)

save('est'1520011)
```

APPENDIX C

Teacher Perception Survey

Answer the questions on the scantron sheet using the following scale:

(A) = Strongly Agree (B) = Agree (C)= Disagree (D) Strongly Disagree

12. I feel supported by my building principal.	**A**	**B**	**C**	**D**
13. I feel supported by the administrative staff.	**A**	**B**	**C**	**D**
14. I feel supported by my professional colleagues.	**A**	**B**	**C**	**D**
15. I believe I have opportunities to grow professionally as I fulfill duties at my ISD.	**A**	**B**	**C**	**D**
16. I believe we spend too much time focusing on standardized tests.	**A**	**B**	**C**	**D**
17. I believe my contributions are appreciated by my colleagues.	**A**	**B**	**C**	**D**
18. I need more support in meeting the needs of my most challenging students.	**A**	**B**	**C**	**D**
19. I believe "all" students in my ISD are treated equitably regardless of race, culture, disability, gender or social economic status.	**A**	**B**	**C**	**D**
20. I believe my ISD families of are supportive of our mission to effectively teach all students.	**A**	**B**	**C**	**D**
21. I believe my ISD families of African American students are supportive of our mission to effectively teach all students.	**A**	**B**	**C**	**D**
22. I believe the district has strong support for academic excellence from our surrounding community (civic, church, business).	**A**	**B**	**C**	**D**
23. I believe some students do not want to learn.	**A**	**B**	**C**	**D**

24. I believe teachers should be held accountable for effectively teaching students who live in adverse circumstances. **A B C D**
25. I believe there are factors beyond the control of teachers that cause student failure. **A B C D**
26. I believe the in-service training this past year assisted me in improving my teaching strategies. **A B C D**
27. I believe I am culturally responsive in my teaching behaviors. **A B C**
28. I believe cooperative learning is an integral part of my ISD teaching and learning philosophy. **A B C D**
29. I develop my lessons based on Texas Essential Knowledge and Skills (TEKS). **A B C D**
30. I believe African American students consider performing well in school as "acting White." **A B C D**
31. I believe African American students have more behavior problems than other students. **A B C D**
32. I believe African American students are not as eager to excel in school as White students. **A B C D**
33. I believe teachers engage in bias behavior in the classroom. **A B C D**
34. I believe students who live in poverty are more difficult to teach. **A B C D**
35. I believe African American students do not bring as many strengths to the classroom as their White peers. **A B C D**
36. I believe students that are referred to special education

usually qualify for special education services in our school. **A B C D**

37. I believe it is important to identify with the racial groups of the students I serve. **A B C D**
38. I believe I would prefer to work with students and parents whose cultures are similar to mine. **A B C D**
39. I believe I am comfortable with people who exhibit values or beliefs different from my own. **A B C D**
40. I believe cultural views of a diverse community should be included in the school's yearly program planning. **A B C D**
41. I believe it is necessary to include on-going family input in program planning. **A B C D**
42. I believe I have experienced difficulty in getting families from African American communities involved in the education of their students. **A B C D**
43. I believe when correcting a child's spoken language, one should model appropriate classroom language without further explanation. **A B C D**
44. I believe there are times when the use of "non-standard" English should be accepted in school. **A B C D**
45. I believe in asking families of diverse cultures how they wish to be identified (e.g., African American, Bi-racial, Mexican). **A B C D**
46. I believe that in a society with as many racial groups as the United States, I would accept the use of ethnic jokes or phrases by students. **A B C D**
47. I believe there are times when "racial statements" should

be ignored. **A B C D**

48. I believe a child should be referred "for testing" if learning difficulties appear to be due to cultural differences. **A B C D**
49. I believe the teaching of ethnic customs and traditions is not the responsibility of public school personnel. **A B C D**
50. I believe Individualized Education Program meetings or planning should be scheduled for the convenience of the family. **A B C D**
51. I believe frequently used material within my class represents at least three different ethnic groups. **A B C D**
52. I believe students from certain ethnic groups appear lazy when it comes to academic engagement. **A B C D**
53. I believe in-service training focuses too much on "multicultural" issues. **A B C D**
54. I believe I address inappropriate classroom behavior even when it could be easily be ignored. **A B C D**
55. I believe I am able to effectively manage students from all racial groups. **A B C D**
56. I believe I have a clear understanding of the issues surrounding classroom management. **A B C D**
57. I believe I have a clear understanding of the issues surrounding discipline. **A B C D**

Table C1
Principal Component Analysis with Varimax Rotation

	Factors							
Item #	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
30	***.513***	.038	.015	-.026	.066	-.115	.014	.134
31	***.785***	.020	.063	.064	-.016	.034	.063	-.045
32	***.810***	.054	.026	.078	.108	.012	.078	.032
34	***.579***	.063	.055	-.007	-.168	.235	.016	.116
35	***.745***	.080	.051	.026	.103	.090	.076	-.104
38	***.444***	.048	.149	.003	.037	.097	.155	-.065
42	***.503***	-.021	-.012	.293	-.051	.039	.022	.136
52	***.557***	.001	.071	.192	.077	-.042	.260	.237
12	.080	***.712***	-.016	.067	.043	.153	-.044	-.152
13	.090	***.765***	.028	.084	-.009	.129	-.046	-.092
14	.042	***.701***	.063	.129	.050	-.101	.062	.090
15	.057	***.652***	-.030	.194	-.013	.208	-.006	-.012
17	.001	***.666***	.123	.187	.040	-.018	.050	.137
55	.179	.036	***.784***	.084	.100	.118	.117	-.089
56	.034	.037	***.911***	-.025	.061	.129	.046	.041
57	.040	.050	***.903***	.016	.058	.112	.023	.040
19	-.013	.288	-.025	***.480***	.054	.078	.047	.001
20	.057	.189	.038	***.775***	-.021	.085	-.017	.080
21	.229	.130	.015	***.804***	.048	.044	-.060	.035
22	.102	.183	.046	***.581***	.003	.152	-.028	-.006
37	.029	.095	.141	-.004	***.472***	.026	-.035	.019
39	.037	-.013	.108	.034	***.612***	.103	.127	-.123
40	.035	.001	.087	-.096	***.626***	.294	.056	.146
41	.067	.028	.064	.013	***.516***	.321	-.018	.136
50	.074	-.028	-.014	.075	***.519***	-.031	.061	-.053
26	.007	.299	-.132	.284	-.039	***.500***	-.026	.002
27	.074	.036	-.303	.076	.132	***.560***	.054	-.019
28	.090	.211	.201	.118	.151	***.598***	.026	.091
51	.068	-.117	.093	.095	.277	***.423***	-.073	-.134
46	.151	.017	.092	-.015	.038	.033	***.704***	.016
47	.160	.008	.012	-.014	-.004	.039	***.719***	-.005
48	.074	-.047	.167	-.048	.075	-.078	***.451***	.388
23	.242	.073	-.017	.117	.123	-.028	-.082	***.489***
25	.149	-.055	-.001	.187	-.154	-.143	-.229	***.493***
49	.111	-.025	-.092	.026	.142	.320	.235	***.523***
53	.210	.070	.042	.034	.154	.101	.264	***.417***
% Var. Explained	7.78%	6.38%	5.81%	5.36%	4.81%	4.06%	3.90%	3.41%

Rotations converged in 60 iterations

Table C2
Reliability Statistics for the Factors

Factor	No. of Items	Cronbach's Alpha
Factor I	8	.805
Factor II	5	.797
Factor III	3	.876
Factor IV	4	.735
Factor V	5	.601
Factor VI	4	.517
Factor VII	3	.527
Factor VIII	4	.463

VITA

Name:	Prathiba Natesan
Address:	2401 South Ocean Drive, Hollywood, FL 33019
Email:	prathibachaj@gmail.com
Education:	B. Arch, University of Madras, 2001
	M.S. Construction Management, Texas A&M University, 2003
	Ph. D. Curriculum and Instruction, Texas A&M University, 2007